Recognizing GSM Digital Speech

Ascensión Gallardo-Antolín, Carmen Peláez-Moreno, *Member, IEEE*, and Fernando Díaz-de-María, *Member, IEEE*

***Abstract*—The Global System for Mobile (GSM) environment encompasses three main problems for automatic speech recognition (ASR) systems: noisy scenarios, source coding distortion, and transmission errors. The first one has already received much attention; however, source coding distortion and transmission errors must be explicitly addressed. In this paper, we propose an alternative front-end for speech recognition over GSM networks. This front-end is specially conceived to be effective against source coding distortion and transmission errors. Specifically, we suggest extracting the recognition feature vectors directly from the encoded speech (i.e., the bitstream) instead of decoding it and subsequently extracting the feature vectors.**

This approach offers two significant advantages. First, the recognition system is only affected by the quantization distortion of the spectral envelope. Thus, we are avoiding the influence of other sources of distortion as a result of the encoding-decoding process. Second, when transmission errors occur, our front-end becomes more effective since it is not affected by errors in bits allocated to the excitation signal. We have considered the half and the full-rate standard codecs and compared the proposed front-end with the conventional approach in two ASR tasks, namely, speaker-independent isolated digit recognition and speaker-independent continuous speech recognition. In general, our approach outperforms the conventional procedure, for a variety of simulated channel conditions. Furthermore, the disparity increases as the network conditions worsen.

***Index Terms*—Coding distortion, Global System for Mobile (GSM) networks, speech coding, speech recognition, tandeming, transmission errors, wireless networks.**

I. Introduction

THE outstanding success of mobile phone communications all over the world has brought up the possibility of creating a new range of services which make use of the inherent ubiquity of mobile facilities. These new phone-enabled services can make the most of the automatic speech recognition (ASR) technology, which provides them with a more natural and easy-to-use interface, even more advantageous if we consider that the usual size of mobile devices is very small.

We are however still far from getting these ASR systems to work properly in this mobile environment. Effective procedures must be developed to tackle the new sources of degradation produced by digital mobile telephony systems. The main ones are as follows.

Noisy scenarios: a mobile communications scenario is inherently noisy, with high and very variable noise levels (many different situations: public places, car cockpit, etc., hands-free operation mode, etc.) [10].

Speech codec (encoder-decoder) distortion: standard codecs are designed to work at specific bit rates while maintaining the perceptual quality as high as possible. However, the distortion introduced by the codec, which becomes higher as the bit rate lowers, can not be ignored [23], [42], [49].

Transmission errors: due to the unreliable nature of the radio-frequency channel, transmission errors are much more influential than over-wired links.

The noisy speech problem has been addressed in different ways in the context of GSM environment: speech enhancement (e.g., [43]), robust parameterizations (e.g., [10]), model compensation (e.g., [49]), etc. While this problem is extremely important, especially if we take into account that the aforementioned ubiquity highly increases the number of physical situations in which a conversation can take place, we have focused our attention on the more specific of the wireless transmission problems: speech coding distortion and transmission errors.

The need to cope with damaging codec effects is a problem peculiar to mobile communications and some other networks like Internet. Due to the scarcity of bandwidth, habitual in those systems, the speech signal should be notably compressed. On doing away with signal redundancy, speech codecs impoverish the acoustic-phonetic information sent over the wireless channel. ASR systems have been shown to be sensitive to this impoverishment [10]. Speech codecs are typically designed to maximize a mean opinion score (MOS) [24]. However, more effective design strategies could be developed for ASR purposes. In this sense, some authors (see, for example [8], [12] or [46]) have pushed a distributed speech recognition (DSR) approach consisting, as we shall review later, of sending the specific parameters for ASR to the receiving end. This approach can be very useful when both the (service) client and server can negotiate the parameters of the transaction or rely on some a priori agreement. Alternatively, as illustrated later, our proposal provides an enhanced recognition relying exclusively on the ASR server capacities, whichever terminal accessing those services is. This allows any of the currently available mobile phones to use any of these speech-enabled services.

Transmission errors also play an important role in digital mobile phone communications. With respect to them, we would highlight that their impact on the speech signal is conceptually quite different to that due to noise. While noise is typically modeled as additive, convolutive or even a combination of both, the influence of transmission errors on the speech signal depends on the speech coding algorithm itself and the bitstream structure.

In this paper, we propose a specifically designed front-end to deal with both, speech coding distortion and transmission errors. In particular, we suggest deriving the ASR parameterization straight from the digital speech representation (i.e., the

Manuscript received November 26, 2002; revised September 6, 2004. This work was supported in part by Spanish Regional Grant CAM-07T-0031-2003 and the Spanish Government under Grant TIC2002-02025. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Tanja Schultz.

The authors are with the Signal Theory and Communications Department, University Carlos III de Madrid, 28911-Leganes (Madrid), Spain (e-mail: gallardo@tsc.uc3m.es; carmen@tsc.uc3m.es; fdiaz@tsc.uc3m.es).

Digital Object Identifier 10.1109/TSA.2005.853210

bitstream) that travels over the wireless channel. The rationale for this proposal is twofold. On the one hand, we have access to the original (though quantized) speech parameters (those extracted by the speech encoder), thereby preventing the encoding-decoding process from influencing the recognizer performance. On the other hand, since we only use a subset of the encoded parameters (those relevant for ASR purposes), transmission errors occurring on other parameters will not decrease the recognition performance. We introduced preliminary encouraging results ([26] and [27]) in parallel with Huerta *et al.* [33]. These two pioneering works were followed by the ones by Choi *et al.* [3] and Kim *et al.* [37]. Furthermore, we have also applied these ideas for ASR over the Internet [45].

Here, we have thoroughly tested the performance of our system in a realistic environment using two different ASR tasks (isolated digit and continuous speech recognition). We have focused on the GSM (Global System for Mobile, European ETSI Standard) system, using the GSM standard codecs, considering tandeming scenarios and modeling the GSM channel; our conclusions are however quite general and could be applied to other cellular systems.

The GSM communication channel has been simulated. The simulation method is based on a combination of available theoretical results, GSM specifications and measured data. Details of the channel simulation procedure are given in the Appendix. Furthermore, GSM channel coding has been analyzed and found to be favorable to our approach, as shall become apparent in Section III.

Besides, a novel procedure to estimate an energy parameter from the well-protected parameters of the bitstream is devised for the "full rate" speech coder and described in Section IV-B3.

This paper is organized as follows. Section II presents alternative architectures for ASR in the GSM environment and the specific problems that should be affronted and more especially, discussing the influence of coding distortion and transmission errors. Section III reviews the main characteristics of the speech coding algorithms chosen for this work. Section IV tidily describes our proposal in comparison with the conventional approach. Section V presents the experiments and discusses the results, highlighting the key issues in ASR in the GSM environment. Finally, some conclusions are drawn and the main areas for future work are outlined in Section VI.

II. Speech Recognition in the GSM Environment

A. Alternative Architectures for ASR in the GSM Environment

Digalakis *et al.* ([8]) distinguished three possible ASR architectures depending on how the speech recognition processing is distributed between the mobile terminal and the machine running the ASR application, i.e., between the client and server sides, namely: 1) local recognition (client-only processing); 2) remote recognition (server-only processing); and 3) distributed recognition (client-server processing).

1) Local Recognition: The best way to avoid both coding distortion and transmission errors is, needless to say, to perform the speech recognition at the user local terminal. In this case, speech coding distortion and transmission errors are not a problem. Nevertheless, this approach has two important disadvantages: first, the application must reside at the local terminal, which must support the whole computational load; and second, it is not possible to reproduce the speech signal at the remote end.

2) Remote Recognition: The best alternative to reduce the computational load at the local terminals is, obviously, to let the server perform all the recognition process. In this case, voice should be transmitted over the mobile network and consequently will be affected by the already mentioned distortions.

In this case, the mobile terminal has to know nothing about the kind of application running at the remote end, the only requirements being the use of a standard codec supported at both ends. This fact becomes relevant for the design of applications that integrate voice, data or any other kind of media, since it allows universal access from almost any terminal.

3) Distributed Recognition: A compromise between the two previous solutions consists on performing part of the recognition process at the client end (namely, the parameterization), and the remaining at the server end (see for example, [8], [46], or [12]). The advantages of this approach rely on the fact that the bandwidth required to transmit the recognition parameters is very small, while the computational effort needed for the parameter extraction is not so high.

With respect to its shortcomings: first, a standardized front-end is needed so that every client terminal computes and transmits the same parameters; second, the impossibility of reproducing or processing the speech signal at the remote end; and third, the increase in the requirements on the terminals with respect to remote recognition.

Although this alternative seems promising (in fact, much effort is currently being made to address its shortcomings; see [50] and [4], for example), we have decided to explore the remote recognizing alternative, since in our opinion, it offers two relevant advantages: 1) it does not impose restrictive conditions on the client terminal's capabilities nor does it create the need for special setting or agreements between client and server and 2) it preserves the transmission bandwidth requirements and the compatibility with the existing standard-based voice applications.

B. Remote Recognition: Key Issues

As briefly stated in the Section I, the GSM environment entails three main problems for ASR systems: noisy scenarios, source coding distortion and transmission errors. Obviously, these three are not the only ones and, although not considered in this paper, it should be mentioned that other subsystems of the GSM system could also affect the performance of a remote ASR system; for example, the discontinuous transmission (DTX) or the insertion of comfort noise. The first one, though conservatively designed, occasionally causes some clipping of the speech signal, while the second inevitably (though slightly) disturbs the estimation, at the remote server, of the background noise characteristics. Furthermore, other parts of the system having to do with the signaling process, such as the handover protocol, can also affect (in our opinion, very sporadically) the performance of an ASR system.

In the following subsections we present the three main above mentioned problems, focusing on the last two (source coding

distortion and transmission errors), which have received much less attention and constitute the motivation of our work.

1) Noisy Scenarios: The inherent mobility entailed by GSM allows the user to make calls from almost anywhere: public places, within a car, at the roadside, etc. As a consequence, reliable ASR systems should be robust to any kind of background noise as well as the Lombard effect.

Current research addressing the noise problem, i.e., to reduce the mismatches between training and testing conditions, advances along three separate directions [30] (some authors divide them only into two classes, merging the first two): first, speech enhancement (e.g., [58] and [43]); second, more robust parameterization techniques impervious to noise effects (e.g., [51], [44], [6] and [10]); and third, techniques to adapt clean-speech Hidden Markov Models to noisy speech conditions (e.g., [25], [41] and [49]). These techniques can be combined to improve [10]. A survey of these techniques can be found in [32] and [34].

2) Influence of Coding Distortion on Speech Recognition: At the typical rates of cellular systems, codecs based on the source-filter model are used most of the times. These codecs achieve their medium or low bit rates by assuming a simplified speech production model with negligible interaction between source and filter. The filter is determined on a frame-by-frame basis while the excitation is computed with a higher time resolution (from two to four times per frame, depending on the codec) usually by means of an analysis-by-synthesis procedure aimed at minimizing a perceptually weighted version of the coding error. As a result, it can be said that these codecs introduce two different types of distortion, namely, the one caused by the quantization of the parameters to be transmitted, and the arising from the inadequacy of the model itself. Consequently, the waveform, short-time spectrum, and other relevant characteristics of the (encoded and) decoded speech signal are somewhat different from those of the original one.

Very limited work has been reported on the influence of source coding distortion on speech recognition. We name three papers that address this problem directly, the first by Euler and Zinke [23], the second by Dufour *et al.* [10], and the third by Lilly and Paliwal [42].

Euler and Zinke analyze the influence of different codecs in the range of 4.8 kb/s to 64 kb/s, drawing the following conclusions: 1) Applying systems trained with 64 kb/s speech to lower rate coded speech substantially increases the error rate of the recognizer. 2) The recognition rates for coded speech significantly improve when the speech recognizer is trained using the same coding algorithm. However, for bit rates below 16 kb/s, they find recognition losses even with matched training and test conditions. 3) Using a Gaussian classifier to guess the encoding method (with a correct classification rate of about 96%), the recognition performance is the same as with known coding conditions.

Lilly and Paliwal [42] conclude that speech codecs with a bit rate of 16 kb/s and above display good recognition performance, even when the speech signal goes through several tandeming stages (successive encoding and decoding processes). However, as the bit rate decreases below 16 kb/s, the loss of recognition accuracy becomes more significant, and even more with tandeming.

Since speech codecs for mobile telephony operate, in general, below 16 kb/s (in particular, the half and full-rate GSM codecs work at 5.6 and 13 kb/s at rates, respectively), it can be concluded that the speech coding distortion will significantly affect the recognition performance.

Dufour *et al.* [10] evaluate a robust parameterization, namely RN_LFCC ('Root-Normalized Linear Frequency Cepstral Coefficients'), in the GSM environment, including the source coding effects. Specifically, they perform several speech recognition experiments with speech coded by the full-rate (FR) and half-rate (HR) standard codecs. From their results it can be concluded that the recognition losses are significant in both cases, but more important in the HR codec.

Besides, the effect of tandeming (addressed by Lilly and Paliwal) deserves, in our opinion, a separate comment for two reasons: first, its influence on recognition performance is very serious; and second, tandeming is very common in practice. The reason for this is that, though once the speech signal has been encoded there is apparently no reason to decode it until it reaches the end user (assuming that all the communication network is digital), this is not the situation in real operation. Actually, when the signal goes through international links, it is usually decoded and re-encoded, using G.711, G.726 (usually at 32 kb/s) [1], or G.728, to undergo the international segment, to be decoded and re-encoded again, now using one of the GSM standards, when it enters the mobile network. On the other hand, even when the signal does not cross political borders, it occasionally suffers the same tandeming process when the near- and far-end telephone operators are different. Finally, for networking reasons, and more frequently than suspected, the speech signal goes through two or more GSM encoding-decoding stages.

Therefore, a realistic evaluation of the influence of speech codecs on remote ASR systems should consider tandemings.

3) Transmission Errors and Lost Frames: Transmission errors inevitably form part of the GSM environment. Thus, these errors should be included in the benchmark experiments.

The GSM system itself, as with every mobile phone system, provides a mechanism to protect the speech signal against transmission errors: the channel coding. More specifically, the channel encoder (which is explicitly designed for each standard codec) classifies the source bits in several categories depending on their relative perceptual impact, as shall be explained in Section III. In this way, not only are some errors detected and even corrected, but their influence on the (decoded) speech perceptual quality is also minimized. Nevertheless, the GSM channel coding is not capable of detecting and correcting all the errors, and some of them may be present in frames labeled as correct.

Furthermore, when the channel decoder considers that a speech frame is seriously damaged (because the most critical parameters are unreliable due to errors), the 'Bad Frame Indicator' (BFI) is triggered (by the channel decoder) and the frame is discarded—not conventionally decoded (reconstructed)—. Instead, the frame is substituted by an attenuated version of the last reliably received one. If various consecutive frames are seriously damaged, the attenuation increases and when the number of replaced frames is more than five (100 ms), the decoder mutes the output, i.e., a sudden disappearance of signal

occurs ('GSM holes' [36]). These holes, as can be imagined, drastically affect the performance of an ASR system.

III. GSM CODECS

With the objective of providing a better understanding of the technique we are proposing, an outline of some relevant features of GSM codecs follows.

The European Telecommunications Standards Institute (ETSI) has standardized two types of speech traffic channels: full-rate (22.8 Kb/s.) and half-rate (11.4 Kb/s.). However, due to the unreliability of the radio transmission channel, a considerable portion of the bit rate capacity is devoted to channel coding.

Phase 1 of GSM standardization, initially described the FR speech codec [17] which made use of 13 Kb/s, leaving the remaining 9.8 Kb/s. to the channel coder. Phase 2 introduced two new codecs: the "Enhanced Full Rate" (EFR) [21] for the full-rate speech channel, and the HR [19] for the half-rate speech channel. The former employs 12.2 Kb/s. and the latter 5.6 Kb/s. Finally, an adaptive multirate (AMR) [22] set of coders has been defined to obtain the best speech quality, dynamically balancing the allocation of bits between the speech and channel coder. Thus, the AMR system adapts to channel conditions by selecting the appropriate mode (full or half-rate) and the speech coding rate that allows sufficient error protection for the actual error level of the channel.

A. Full-Rate Standard at 13 kb/s

The GSM FR speech codec operates at 13 kbits/s and uses a regular pulse excited linear predictive coder with a long term predictor loop (RPE-LTP) [54]. Basically, the input speech is split into 20 ms-long frames, and for each frame a set of 8 short-term linear prediction (LP) coefficients are found. These LP coefficients are coded as Log Area Ratios (LAR) with 36 bits, more bits being assigned to the lower coefficients. It is worth noting (we will get back to this issue later) that the number of LP coefficients chosen for this standard is significantly lower than for the rest of the GSM codec standards.

Then, each frame is further split into four 5-ms subframes, and for each subframe the encoder finds a delay and a gain for the long-term predictor. Finally, the residual signal after both short and long term filtering are quantized for each subframe will be as follows. The 40-sample residual signal is decimated into three possible excitation sequences, each 13 samples long. The sequence with the highest energy is chosen as the best representation of the excitation sequence. It is encoded normalized by the transmitted parameter $X_{\max}$, using an indicator of the RPE position that identifies the chosen sequence among those possible and finally, each sample in the sequence is quantized with three bits. At the decoder, the reconstructed excitation signal is fed through the long term and then the short-term synthesis filters to give the reconstructed speech. Finally, a postfilter is used to improve the perceptual quality of the reconstructed speech [17].

The encoded speech transmitted over the radio interface must be protected from errors. GSM uses convolutional encoding and block interleaving to achieve this protection. From subjective testing, it was found that some bits were more important for perceived speech quality than others. Thus, in total a FR frame consisting of 260 bits divided into three classes:

- Class Ia: the 50 bits most sensitive to errors.
- Class Ib: 132 bits classified as moderately sensitive to errors.
- Class II: 78 bits classified as least sensitive to errors.

Channel coding details can be found in [14]. However, it is worth noting that most of the bits required to perform the proposed bit-stream based recognition that will be presented in Section IV-B are in Class Ia and therefore among the most protected bits of the transmission.

As regards the procedures to mitigate the effects of damaged frames, the GSM standard does not specify which one should be used [18]. In this paper we have used the most widely implemented one, i.e., repeating the last LP valid filter (maybe with some formant bandwidth expansion), and the synthesis of an excitation based on the last correctly received one.

B. Half-Rate Standard at 5.6 kb/s

The HR codec uses the vector sum excited linear prediction (VSELP) paradigm [29]. A tenth-order LP analysis is performed over a 20 ms-long speech frame. Those coefficients are subsequently converted into ten reflection coefficients (RC) and quantized. Thus, the spectral information here, is more accurate than in the FR codec, which only considers 8 coefficients, and also includes a procedure of soft linear interpolation of the LP coefficients. This interpolation generates a smooth transition between LP coefficients of consecutive frames and can be turned on and off to achieve a better prediction gain in a particular frame. Therefore, it is a frame-by-frame decision made by the coder, which sends a flag to the decoder to indicate it [19].

Four voicing modes are considered, selected using a combination of open loop and closed loop techniques for long term prediction. The excitation codebook consists of 2^{M} code-vectors which are constructed from M basis vectors.

Channel coding is analogous to the one described for the FR codec ([14]). Here, the frame consists of 112 encoded bits split into the following corresponding three classes.

- Class Ia: The 22 bits most sensitive to errors.
- Class Ib: 73 bits classified as moderately sensitive to errors.
- Class II: 17 bits classified as least sensitive to errors.

Again, as with the HR codec, the distribution of the bits into these classes benefits the proposed digital front-end (see Section IV-B) since most of the ones required for this parameterization are among the most protected ones.

The procedure for recovering from errors relies on the two types of indicators that determine the importance of the errors detected: the bad frame indication (BFI) and the unreliable frame indication (UFI). If the BFI flag is set, the speech decoder performs frame substitution and muting (when several consecutive erroneous frames are received). If the UFI flag is set, the speech decoder performs a plausibility analysis of the received parameters and of the output signal. If the frame is considered to be usable, some properties of this output signal are compared to those of the previously valid ones. In the case

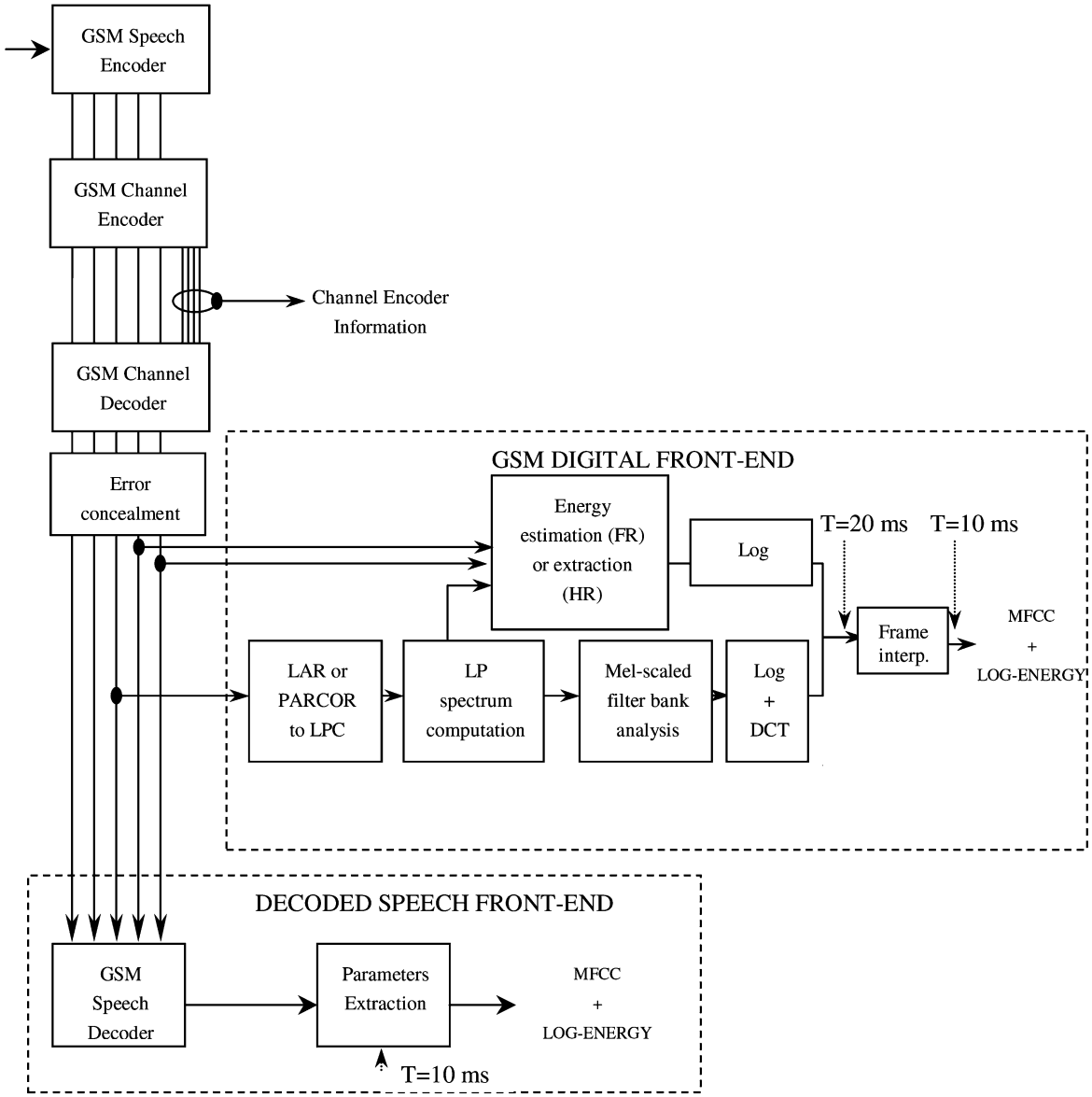

Fig. 1. Parameterization procedures. The lower part of this block diagram illustrates the steps followed in the conventional approach ("decoded speech front-end"), i.e., the encoded speech is received at the far end and subsequently decoded before being parameterized for recognition. The upper part of the diagram represents our proposed procedure ("GSM digital front-end"), where no decoding is performed. Instead, the parameterization is extracted from the quantized spectral coefficients transmitted by the codec.

of large differences, the output signal can be modified to limit them [20].

IV. Recognition From GSM Digital Speech

The essential difference between a conventional ASR system and our approach is the source from which the feature vectors are derived. Thus, to assess our proposal, we have compared the two ASR systems that can be observed in Fig. 1.

The decoded speech-based front-end—from now on denoted "decoded speech front-end"—starts from the decoded speech and proceeds as a conventional ASR system. On the contrary, the GSM digital speech-based one —from now on "GSM digital front-end"—starts from a (quantized) LP spectrum plus a reduced set of parameters extracted from the GSM bitstream. These two different ways of computing the feature vectors are more deeply described in the next subsections.

A. Decoded Speech Front-End

In this conventional approach the feature extraction is carried out on the decoded speech signal, which is analyzed once every 10 ms employing a 25 ms analysis Hamming window, using the HTK package [57]. Twelve Mel-frequency cepstral coefficients (MFCC) are obtained using a mel-scaled filterbank with 40 channels. Then, the log-energy, the twelve delta-cepstral coefficients and the delta-log energy are appended, making a total vector dimension of 26.

B. GSM Digital Front-End

1) Motivation: Standard speech codecs are completely (bit-level) defined. Therefore, it is possible to selectively access the relevant parameters (from the recognition point of view). The underlying idea here is to feed the speech recognizer with a

parameterization directly derived from the GSM digital speech representation, i.e., recognizing from the bitstream. This is feasible because, fortunately, most of the standard codecs used in digital mobile telephony are LP-based, and this type of codecs extract and code the appropriate spectral information as separated parameters, from which recognition can be successfully carried out.

One of the aims of our proposal is to reduce the influence of coding distortion on ASR systems performance. Specifically, the spectral envelope derived from the bitstream is the same that would have been obtained from the original speech, except for the quantization. But, as revealed by the work by Tucker *et al.* [53] and confirmed by our experimental results, the quantization distortion of the parameters that encode this spectral envelope, does not especially affect the recognition performance. On the other hand, the spectral envelope estimated from the decoded speech could exhibit important differences with respect to the original one, since, as will be explained in detail in Subsection IV-B2, the decoded speech is affected by both the quantization distortion of every parameter involved in the speech synthesis (and not only by the quantization of the spectral information ones) and the inadequacies of the source-filter model.

Furthermore, when dealing with *residual transmission errors* (those not detected or corrected by the channel decoder) and *frame erasures*, our front-end turns out to be more effective and robust than the conventional one due to the following reasons: on the one hand, and with respect to residual errors, only those affecting the spectral envelope and energy encoding will damage our system performance, while the remaining (those affecting any bit representing the excitation) will not. On the other hand, and regarding frame erasures, our front-end avoids being affected by the excitation-based part of the decoder concealment procedures. This is not the case for the conventional decoded speech-based approach, since it estimates the spectral envelope from the decoded speech, which exhibits degradations due to the effects of these erroneous information based concealment procedures on both spectral envelope and excitation.

Specifically, for the FR codec, our approach uses all of the bits belonging to class Ia (the most protected ones), 22% of bits belonging to class Ib (moderately protected) and 22% of bits belonging to class II (not protected). For the HR codec, our approach uses 55% or 59% (depending on the voice mode) of the bits belonging to class Ia, 29% or 27% (depending on the voice mode) of the bits belonging to class Ib and none of the bits belonging to class II. By contrast, in both cases, the decoded speech front-end uses the whole bitstream for the decoding process.

Summing up, the advantages of the proposed approach are as follows.

1) The performance of our system is only affected by the quantization distortion of the spectral envelope and the reduced subset of the excitation parameters from which we extract the energy information (see Subsection IV-B3 for the description of the estimation procedure). Thus, we are avoiding the distortions due to the quantization of the remaining parameters and possible inadequacies of the source-filter model.
2) When residual errors or frame erasures occur, our front-end can be more effective since it is practically immune to errors affecting the excitation encoding.
3) The computational effort required is not increased, since the cost of computing the MFCCs from the digital speech is practically equivalent to that of the same task in the conventional front-end; furthermore, in our case, the complete decoding of the speech signal is not necessary.

However, it should be noted that our approach requires the front-end to be adapted to the specific codec the network is using. If not, we would be willing to accept some mismatch. Besides, as we will discuss further on, the spectral envelope is available at the frame rate of the codec (which can be too slow for our purposes). The latter is a minor problem that can be easily solved as will be shown later.

2) Implementation Details: The block diagram in Fig. 1 illustrates the proposed parameterization procedure compared to the conventional one. Our implementation mixes our own procedures with some facilities of the HTK Toolkit. More precisely, the trans-parameterization (from quantized LP parameters to MFCC) is described step by step as follows.

1) For each GSM frame (20 ms of speech for both the half- and the full-rate standards), the P available quantized spectral parameters are extracted from the bitstream. When the codec we are using is the HR, we get ten $(P = 10)$ reflection coefficients and, in the case of the FR, we obtain eight $(P = 8)$ Log Area Ratio (LAR). After having let the standard concealment procedures do their job as recommended in the GSM standard (see Section III), they are converted into P LP coefficients we shall denote a_r, with $r = 1, \ldots, P$ (see [39]).

$$H[k] = H\left(\Omega = k\frac{2\pi}{N}\right) = \frac{1}{1 - \sum_{r=1}^{P} a_r e^{-j\frac{2\pi}{N}kr}} \quad N = 512, \quad k = 0, \ldots, 255 \tag{1}$$

2) Next, a 512-point spectral envelope, $H(\Omega)$, of the speech frame is computed from the P LP coefficients (from which we only use the 256-sample positive half of the full spectrum):
3) A filterbank composed of M mel-scale (we have employed $M = 40$) symmetrical triangular bands identical to the one employed in the conventional front-end is applied to weight $|H[k]|$, yielding 40 coefficients, which are subsequently converted into 12 mel cepstrum coefficients using a discrete cosine transform (DCT) of their log-scaled magnitudes.
4) The frame energy is extracted from the bitstream for the HR codec and estimated for the FR using a small set of related parameters as described in Section IV-B3. The log-energy is appended to the feature vector.
5) A band-limited interpolation FIR filter is applied over the time feature sequences with the purpose of reducing the frame period (the time interval between two consecutive feature vectors) from 20 ms—the one provided by the speech codecs—to 10 ms—the one employed by

the conventional front-end—as it has been observed that recognition figures show a critical dependency on the frame period. This interpolation filter uses the nearest 4 (2 of each side) feature vectors. It is important to note, however, that this filter does not cause any additional delay, since we already admit this delay for the computation of the dynamic parameters.

6) Finally, dynamic parameters are computed (using HTK) for all the 12 MFCC and the log-energy, making a total vector dimension of 26.

3) Energy Estimation: Almost every speech recognizer includes an energy parameter in the feature vector. However, speech codecs do not always explicitly encode this energy as a separate parameter. In any case, it is implicitly encoded in some other parameters, since it can be extracted by decoding the whole speech signal. Nevertheless, in this paper, we advocate for the avoidance of the decoding stage as a means of improving the robustness of the recognizer. This makes it necessary to find a way of obtaining the energy straight from the bitstream by decoding the minimum quantity of information possible.

In [45] we already introduced a procedure to estimate this energy when the codec was the standard G.723.1. In this paper we have made use of two GSM codecs. Fortunately, the HR codec sends the energy information as a separate parameter. However, the FR codec does not perform this explicit encoding and therefore the energy should be estimated.

For that purpose, we have obtained a rough estimate of the mean power of every subframe using the source-filter model in which the codec is based. Thus, the mean power of the synthesized speech, $E\{x^2[n]\}$, in that subframe can be computed as follows:

$$P_x = E\{x^2[n]\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{xx}(\Omega)d\Omega \tag{2}$$

where $\Phi_{xx}(\Omega)$ is the power density spectrum of the synthesized speech.

Modeling the excitation, $e[n]$, as a zero-mean white Gaussian noise, $\Phi_{xx}(\Omega)$ can be expressed as

$$\Phi_{xx}(\Omega) = \sigma_e^2|H(\Omega)|^2 \tag{3}$$

where σ_e^2 is the power spectral density of the excitation and $H(\Omega)$ is the frequency response of the synthesis filter. Introducing (3) into (2) we obtain

$$P_x = E\{x^2[n]\} = \frac{\sigma_e^2}{2\pi}\int_{-\pi}^{\pi} |H(\Omega)|^2 d\Omega. \tag{4}$$

Let $\hat{P}_x[k,i]$ denote the estimated mean power of the subframe i $(0 \leq i < N_{\text{sf}} - 1)$ of the frame k, where N_{sf} is the number of subframes that make up a frame (which is 4 for the FR codec). Following (4), $\hat{P}_x[k,i]$ can be calculated as

$$\hat{P}_x[k,i] = \hat{\sigma}_e^2[k,i] \cdot \hat{E}_h[k,i] \tag{5}$$

where $\hat{\sigma}_e^2[k,i]$ and $\hat{E}_h[k,i]$ represent the estimations of the excitation mean power and the energy of the impulse response of the synthesis filter, respectively. In the following exposition, the frame and subframe indexes, k and i, will be dropped for simplification and recalled appropriately when necessary.

Starting with the filter energy, $\hat{E}_h$, it is easily shown that it can be obtained approximating the integral of the (4) by the following sum involving the 256-point spectral envelope calculated from the LP coefficients, which have previously been computed from the LAR parameters as a part of our parameterization procedure (see previous section):

$$\hat{E}_h = \frac{2}{N}\sum_{r=0}^{\frac{N}{2}-1} \left| H\left(\frac{2\pi}{N}r\right)\right|^2 \tag{6}$$

where $N = 512$ (as we have already mentioned, only the positive part of the spectral envelope is considered).

Another possibility for the estimation of $\hat{E}_h$ without making use of the spectral envelope calculation is to employ the PARCOR (PARtial CORrelation) parameters, p_r (usually obtained as part of the speech decoding process for stability checking—see [39]—), according to the following expression [47]:

$$\hat{E}_h = \frac{1}{\prod_{r=1}^{M}(1-p_r^2)} \tag{7}$$

where M is the number of PARCOR parameters computed by the codec, which is 8 for the FR standard.

Before describing how to estimate the mean power of the excitation, $\hat{\sigma}_e^2$, it is convenient to review the procedure the FR codec employs for its computation. As is common in this family of codecs, the excitation, $e[n]$, can be decomposed into an adaptive contribution, $u[n]$, which captures the periodicity of the signal, and a stochastic one, $v[n]$, aiming at reflecting its randomness, so that

$$e[n] = u[n] + v[n]. \tag{8}$$

At this point, we will assume that these two excitation components are uncorrelated, and therefore

$$\hat{\sigma}_e^2 = \hat{\sigma}_u^2 + \hat{\sigma}_v^2 \tag{9}$$

where $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ are, respectively, the estimated mean powers of the adaptive and stochastic contributions.

In the FR standard codec, the adaptive contribution is implemented as a first-order long-term linear predictor

$$u[n] = \beta_i e[n - L_i], \quad 0 \leq n \leq L_{\text{sf}} - 1, \quad 40 < L_i \leq 120 \tag{10}$$

where the parameters β_i and L_i represent, respectively, the prediction gain and pitch lag for the subframe i. L_{sf} denotes the length (in samples) of the codec subframes ($L_{\text{sf}} = 40$ for the FR codec).

Thus, from (10), $\hat{\sigma}_u^2$ could be computed as follows:

$$\hat{\sigma}_u^2[k,i] = \frac{1}{L_{\text{sf}}}\sum_{n=0}^{L_{\text{sf}}-1} (\beta_i e[n - L_i])^2. \tag{11}$$

However, as can be observed, proceeding in this way would require the whole reconstruction of the excitation signal, $e[n]$.

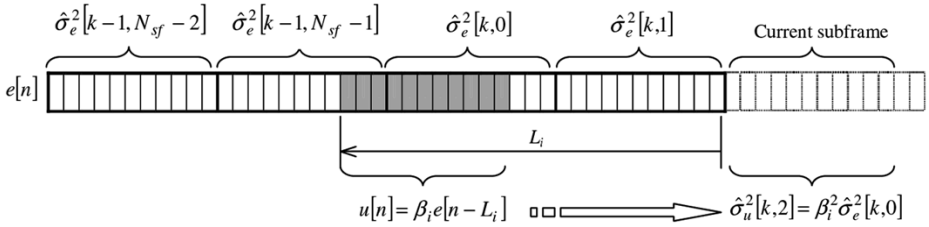

Fig. 2. Illustration of the estimation process of the adaptive contribution mean power. It can be seen how the excitation mean power of the (past) subframe in which the larger portion of this past excitation ($e[n-L_i]$) is, becomes a good choice for the adaptive contribution mean power.

On the contrary, we maintain that the ASR front-end will be robust by avoiding the speech decoding process. Consequently, we suggest roughly approximate $\hat{\sigma}_u^2[k,i]$ as follows:

$$\hat{\sigma}_u^2[k,i] \approx \begin{cases} \beta_i^2\hat{\sigma}_e^2[k,l], & 0 \le l < i \\ \beta_i^2\hat{\sigma}_e^2[k-1,l+N_{\rm sf}], & -N_{\rm sf} \le l < 0 \end{cases} \quad (12)$$

where l can be calculated as

$$l = \left\lfloor i + \frac{1}{2} - \frac{L_i}{L_{\rm sf}} \right\rfloor \quad (13)$$

where $\lfloor x \rfloor$ denotes the biggest integer less or equal to x. This approximation takes into account that, essentially, the adaptive contribution is simply the portion of the past excitation—conveniently weighted by β_i—that achieves the best matching of the signal periodicity (see (10)). However, $e[n-L_i]$ is not available for our purposes. Instead, by the time $\hat{\sigma}_u^2[k,i]$ is requested, we have already computed every previous $\hat{\sigma}_e^2[k,i]$. Therefore, a good choice for this adaptive excitation mean power is the total excitation mean power of the past subframe where the larger portion of this past excitation, $e[n-L_i]$, is (see Fig. 2).

On the other hand, for the estimation of the mean power of the stochastic contribution, we have taken advantage of the fact that besides the information regarding the particular sequence of excitation pulses, an overall gain for each subframe, $X_{\max_i}$, is encoded and transmitted to the receiving end.

Assuming that the overall gain of the corresponding subframe $X_{\max_i}$ may be thought as being proportional to the estimated mean power of the stochastic excitation, $\hat{\sigma}_v[k,i]$, i.e.,

$$X_{\max_i} = K\hat{\sigma}_v[k,i]. \quad (14)$$

We can finally obtain the desired estimation $\hat{\sigma}_v^2[k,i]$ as follows:

$$\hat{\sigma}_v^2[k,i] = \left(\frac{X_{\max_i}}{K}\right)^2 \quad (15)$$

with K being some proportionality constant.

K is empirically estimated from the training speech data. For each subframe i, the mean power of the stochastic part of the excitation, $(\hat{\sigma}_v^2[i])_{\rm train}$, is computed and the value of the overall gain corresponding to this subframe $(X_{\max_i})_{\rm train}$ is extracted from the bitstream. Using (14), the specific value of the constant K_i is calculated for the current subframe. Finally, K is estimated by averaging the values of K_i over all the subframes. Following this procedure, a value of $K = 1.88$ is obtained.

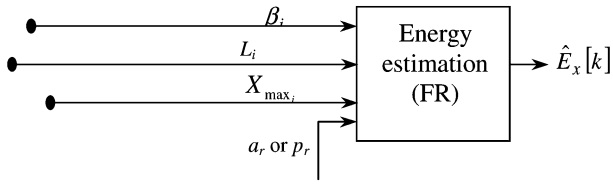

Fig. 3. Parameters involved in the estimation of the energy of the kth frame in the GSM digital front-end (enlargement of Fig. 1), where $0 \le i < 4$ and $1 \le r \le 8$. These parameters are among the most protected by the channel encoder, yielding an enhanced robustness.

Finally, an estimated mean power for the whole frame, $\hat{P}_x[k]$, is computed by averaging $\hat{P}_x[k,i]$ over the $N_{\rm sf}$ subframes. The energy is then calculated as $\hat{E}_x[k] = N_{\rm sf}L_{\rm sf}\hat{P}_x[k]$.

It is important to realize, that for fulfilling σ_x^2 estimation, we have additionally extracted and decoded three parameters for every subframe, namely: the β_i, L_i, and $X_{\max_i}$ (see Fig. 3).

This procedure leads to a more robust ASR parameterization for two fundamental reasons: first of all, the fewer the number of parameters employed, the lower the probability of error in any of these parameters. The inclusion of any of the remaining parameters (which altogether represent 63% of the total number of bits) into the energy computation can only contribute to a worsening of the estimation since the energy information content of these parameters is low and they are not reliable. Second, the channel codec puts more emphasis on protecting the most perceptually relevant parameters. Certainly, the energy is among them and the choice of the β_i, L_i, and $X_{\max_i}$ parameters together with the spectral envelope descriptors is supported by the fact that 52% of the bits that quantify these parameters belong to class Ia (highest protection), 30% to class Ib (medium protection) and the remaining 18% to class II (no protection). By contrast, the rest of the parameters bits belong on a 63% rate to class Ib and 37% to class II.

In Fig. 4 the energy obtained by decoding a whole sentence is depicted against the one resulting from the procedure we propose in the absence of transmission errors. It illustrates how this energy estimate is appropriate for recognition purposes.

4) Related Works: As mentioned in the Introduction, recently some authors have proposed speech recognition systems following the same fundamental idea: recognizing from the bitstream ([33], [3] and [37]). In this section, we compare our proposed front-end with those described by other authors, highlighting and discussing the differences.

The research of Huerta *et al.* ([33]) focuses strictly on the GSM FR standard. Their experiments reveal that the residual signal (i.e., the excitation of the source-filter model embedded

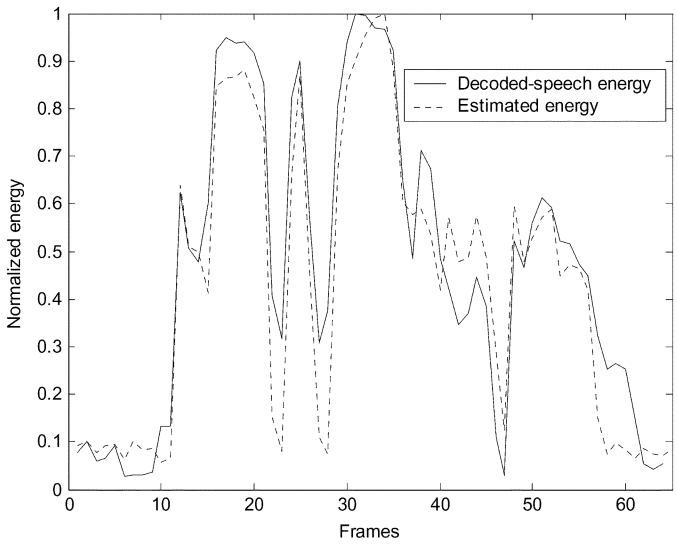

Fig. 4. Energy parameter employed by the conventional front-end (solid line) and the one obtained using the proposed estimation procedure (dashed line) are depicted for the sentence "Clear the data screen" in absence of transmission errors.

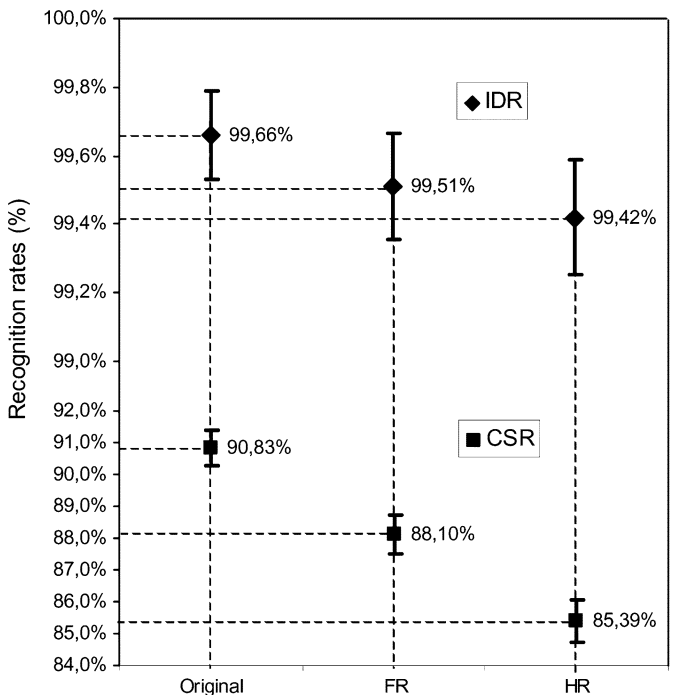

Fig. 5. Illustration of the influence of speech coding on the IDR (upper part of the figure) and CSR (lower part) tasks. For both tasks, recognition rates for original (unencoded) and HR and FR coded speech are presented. The confidence intervals are depicted. As it can be seen, the influence of coding distortion is practically negligible for the IDR task, while it is noticeable for the CSR task.

in the codec) still contains some information relevant to recognition. Therefore, they suggest combining the cepstra derived from LAR with those derived from the residue. Using this method they report a recognition accuracy equivalent to that achieved for the original (unencoded) speech. In addition to the problem of coding distortion, these authors have addressed the problem of additive noise. However, transmission errors and frame erasures are not considered. As we have already stated, one of the strengths of our front-end is its robustness against these last two distortions. Moreover, this robustness originates partly, from the fact that, for every single parameter available from the bitstream, we balance the possible benefits of including relevant information and the risk of getting it corrupted. As much of the information embedded in the residue is not very relevant for ASR and the risk of corruption is higher (due to the weaker protection of the channel codec), the approach we propose deems it more convenient not to include it.

In [3], Choi *et al.*, have used the QCELP (Qualcomm Inc., 1993) to test a bitstream based speech recognizer. This codec represents the spectral envelope through ten LSP. They experimentally show that the quantized LSPs obtained from the bitstream are closer to the original ones than those obtained from the decoded speech. Their main contribution, in our opinion, is to obtain an extraordinarily simple way of transforming the LSP into what they call pseudo-cepstrum (a good approximation to LPC-cepstrum). A significant improvement with respect to the conventional procedure is reported. Nevertheless, they consider neither transmission errors nor frame erasures, which leaves the question of how robust this approximation is in such cases.

The research of Kim *et al.* [37] focuses on the American IS-136 Communication System and uses the IS-641 speech codec [31]. The authors use LP-cepstral coefficients as recognition parameters and suggest including some voiced/unvoiced information in the feature vector. An energy parameter completes their feature vector. Unlike our proposal, the residual signal must be decoded to obtain this last parameter. In this case, the authors consider both channel impairments and car and babble noises. However, with respect to channel impairments, they do not consider transmission errors but concentrate only on frame erasures. Kim *et al.* consider the channel codec powerful enough to deal with residual errors as far as the speech recognizer is concerned. Consequently, their work focused on frame erasures and proposed two techniques to deal with them: an extrapolation algorithm and the deletion of the erased frames based on missing feature techniques. This yields a superior performance over the conventional system. However, in our proposal performance testing, we have considered both individual transmission errors and frame erasures, since we have encountered several of such errors, affecting the recognition performance, not repaired by the GSM channel decoder, especially the information that encodes the residual. This leads us to introduce the energy estimation technique described in the previous section instead of using the one based on the decoded residue. Thus, to avoid the influence of transmission errors affecting the residue, we estimate the energy from a reduced set of parameters, which indeed, happen to be among the most protected by the channel encodec.

Our research, unlike those reported in [3] and [37], focuses on the European GSM system. We have successfully tested our system against the most common MFCC-based front-end and have considered not only transmission errors and frames erasures, but also several network configurations including tandeming encodings.

V. Experimental Results

In this section, we first present the baseline recognition systems, the experiments and databases. Second, we evaluate the influence of coding distortion and transmission errors on a conventional ASR system, keeping the results for reference. And third, we compare the proposed front-end with the conventional one for several practical scenarios.

A. Baseline Systems and Databases

In order to compare the proposed front-end with the conventional one, we have chosen two different tasks: speaker-independent isolated digit recognition (IDR task) and speaker-independent continuous speech recognition (CSR task).

With the objective of stating the statistical significance of the experimental results shown in the next subsections, we have calculated the confidence intervals (for a confidence of 95%) using the following formula ([55, pp. 407–408)]:

$$\frac{\text{band}}{2} = 1.96\sqrt{\frac{p(100-p)}{n}} \tag{16}$$

where p is the word accuracy for the IDR or the CSR tasks and n is the number of examples to be recognized (7920 and 10288 words for the IDR and CSR tasks, respectively). Thus, any recognition rate in the tables below is presented as belonging to the band $[p-(\text{band}/2), p+(\text{band}/2)]$ with a confidence of 95%.

1) Speaker-Independent Isolated Digit Recognition: For the speaker-independent IDR experiments, we have used a proprietary database consisting of 72 speakers and 11 utterances per speaker for the 10 Spanish digits. This makes a total of 7920 audio files. This database was recorded at 8 kHz and in clean conditions. In addition, we have encoded this database using both the FR and the HR GSM standards [17], [19], so that we have three different databases at our disposal.

Since the database is quite limited for achieving reliable speaker-independent results, we have used a k-fold cross validation ([7], Ch. 10.6.4). The basic idea is to split the whole database into k disjoint subsets. The classifier is designed based on the union of k-1 subsets and the remaining subset is used for testing. This process is repeated k times. The advantage of this approach is that all the observations in the database are used for training, enabling their efficient exploitation. In addition, in each trial, testing samples are not used during the training of the classifier, so an unbiased estimate of the performance of the classifier is obtained. In our experimentation, the database is divided into 9 disjoint subsets, each of which comprises 8 speakers (each speaker is not included in more than one subset). We have taken one of these subsets for testing (880 files), leaving the remaining (7 040 audio files) for HMM training. We have repeated this process nine times, so the recognition rates presented in this paper for the IDR task correspond to the average results of these 9 experiments.

The baseline is an isolated-word, speaker-independent HMM-based ASR system developed using the HTK package. Left-to-right HMM with continuous observation densities are used. Each of the whole-digit models contains a different number of states (which is three times the number of phonemes in the phonetic transcription of each digit) and three Gaussian mixtures per state.

2) Speaker-Independent Continuous Speech Recognition: The database that we have used in our speaker-independent continuous speech recognition experiments is the well-known Resource Management RM1 Database [52], which has a 991 word vocabulary. The speaker-independent training corpus consists of 3 990 sentences produced by 109 speakers and the test set contains 1 200 sentences from 40 different speakers, which corresponds to the compilation of the first four official test sets. Originally, RM1 was recorded at 16 kHz and in clean conditions; however, our experiments were performed using a (downsampled) version at 8 kHz. As in the previous section, we have encoded this database using both the FR and the HR GSM standards.

We have employed context-dependent acoustic models, namely: three-mixture cross-word triphones. The synthesis of unseen triphones in the training set was performed through a decision tree method of state clustering. The standard word-pair grammar was used as the language model.

B. Influence of Coding Distortion on Speech Recognition Performance

In this subsection we evaluate the influence of the HR and FR GSM standards codecs on the baseline system recognition performance. In first place, we consider each codec separately, assuming only one coding stage between the mobile terminal and the ASR system. After that, several scenarios involving two and three coding stages (tandeming) are considered.

1) One Coding Stage: We have evaluated the influence of the two considered GSM standard codecs on the two previously described recognition tasks.

Fig. 5 shows our results for the IDR (upper part) and the CSR (lower part) tasks. The results achieved in the reference experiment using original (unencoded) speech are compared with those obtained using the FR and HR codecs in matched conditions (training and testing using decoded speech).

With respect to the IDR task, it may be concluded that none of the standards affects the recognition performance significantly. The HR seems to cause a slightly higher distortion, which is almost statistically significant.

The results are not so optimistic for the CSR task: the use of any of the codecs significantly reduces the recognition performance. The word recognition rate declines from 90.83% to 88.10% (i.e., 3.01% relative degradation with respect to the reference value) for FR and to 85.39% (i.e., 5.99% relative degradation) for HR. It may therefore be inferred that both codecs, and especially the HR, clearly reduce the recognition rate.

Summing up, the adverse influence of GSM standard codecs is very significant for a CSR task, while practically irrelevant for a simple IDR task. On the other hand, the HR codec impoverishes more notably the recognition performance than the FR (as expected, it their bit rates are taken into account).

2) Tandeming: A realistic evaluation of the influence of speech codecs on remote ASR systems should consider the (very likely) tandeming encodings. We suggest several possible tandeming configurations involving the HR and FR GSM standards and the ITU G.726 (operating at 32 kb/s). The selected

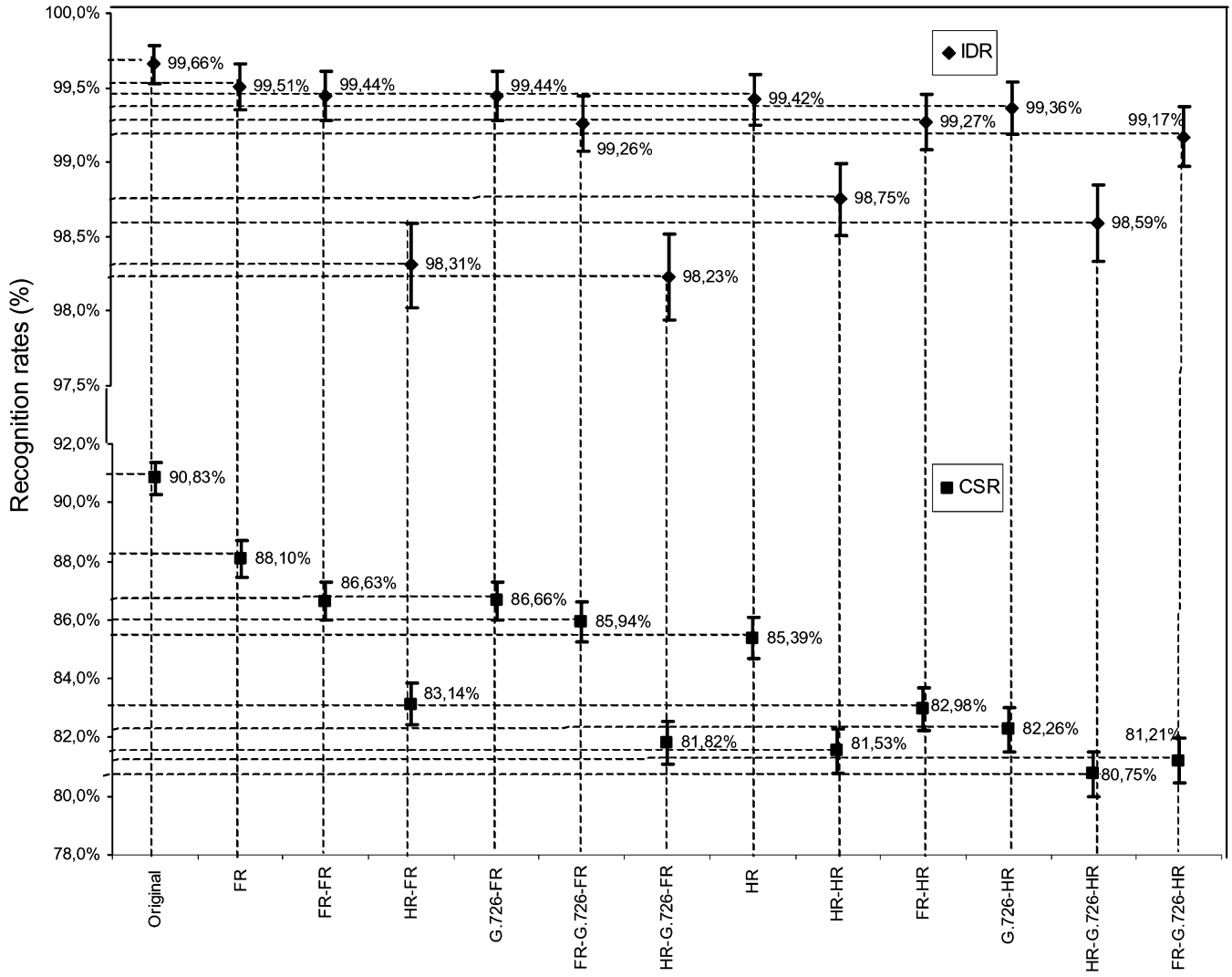

Fig. 6. Illustration of the influence of several stages of coding (tandeming) on the IDR and CSR tasks. Several tandeming configurations are considered involving the HR, the FR and the G.726 at 32 kb/s. The confidence intervals are depicted. As it can be observed, unlike the case of the one coding stage, the influence of coding distortion is significant in most of the cases for both tasks.

tandeming configurations are inspired from three interesting works dealing with this subject: [49], [2], and [9].

Fig. 6 illustrates the influence of tandeming on our IDR (upper part) and CSR (lower part) tasks. When only one coding stage was considered, the influence of the codecs on the IDR task was negligible. Now, even for this simple task, the recognition losses become significant in most cases, especially those involving a HR stage. It is also interesting to note that the G.726 stage has almost no influence on the recognition performance for this task.

In the CSR task, all of the tandeming configurations produce significant losses ranging from 4.6% (for G.726-FR) to 11.1% (for HR-G.726-HR). Again, the impoverishment of the recognition figures is more important when the HR is involved. And, interestingly, the G.726 stage now has a relevant impact on recognition performances.

In sum, if the speech signal reaches the ASR system through several coding stages the influence of the coding distortion is very significant (recognition losses up to 11.1%). Even for the simplest recognition task considered, the losses are significant in most cases. Furthermore, when the task is more complex, even a high-rate coding stage (e.g., G.726 at 32 kb/s) notably affects the performance.

C. Influence of Transmission Errors on Speech Recognition Performance

In this subsection we evaluate the influence of transmission errors on the baseline system recognition performance. We test the recognition system in four or five (depending on the experiment) channel conditions, corresponding to theoretical Bit Error Rates (BER) ranging from 10^{-4} to 5.10^{-2}. The details regarding channel simulation and the characteristics of each of the simulated channels are in the Appendix . In this subsection and the following ones, we refer to each of these channels as its theoretical BER.

Although conveniently explained in the Appendix, it is interesting to highlight that both residual transmission errors (those still present after channel decoding) and frame erasures are jointly considered, since it is the channel decoder which decides whether the frame is discarded (and substituted) or

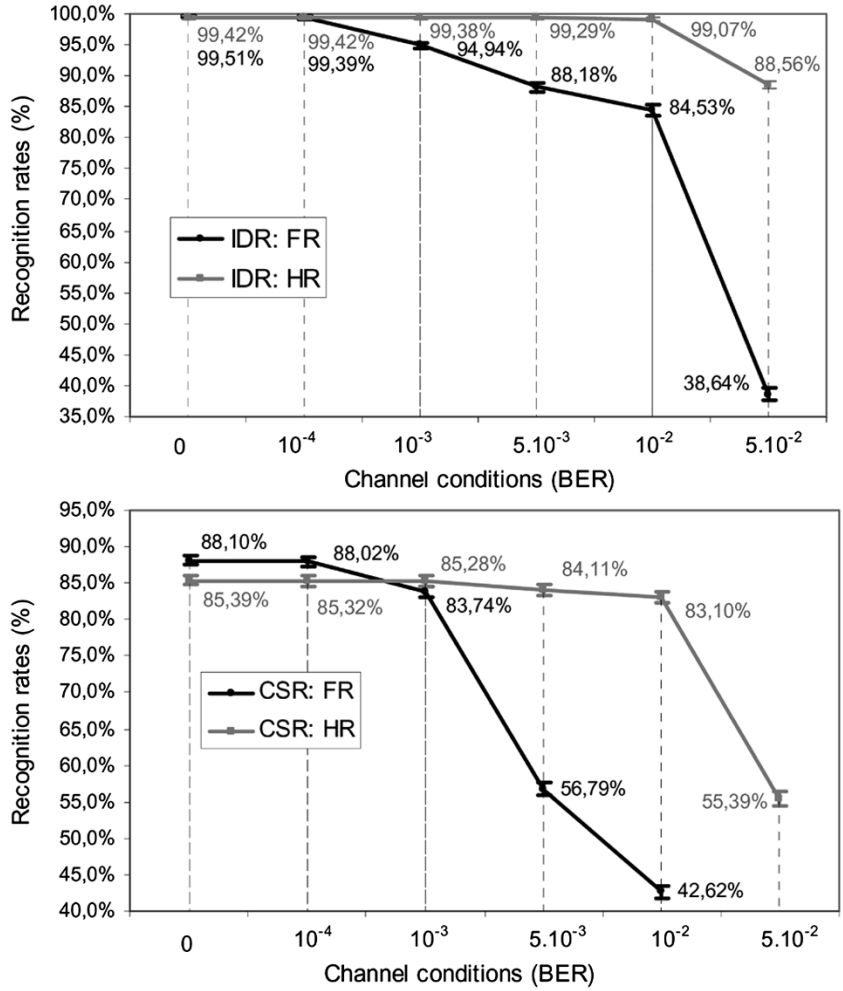

Fig. 7. Influence of transmission errors on the recognition performance for the IDR and CSR tasks. Several channel conditions are considered, corresponding to theoretical BERs ranging from 10^{-4} to 5.10^{-2}. The FR standard turns out to be much more sensitive to transmission errors than HR.

not, depending on the number of bit errors and the class of the affected bits. Thus, once the bitstream has passed through the channel decoder, the source decoder receives a clean frame, a frame with residual errors, or a bad frame indication. In the last case, the source decoder triggers the corresponding frame concealing mechanism. We have included this concealment procedure into both the digital and decoded approaches, with the aim of comparing our system with the best possible opponent.

As we did in the last subsection, we first consider each codec separately, and after, several scenarios involving two coding stages.

1) One Coding Stage: Fig. 7 shows our results for the IDR and CSR tasks. The results depicted for a zero theoretical BER are the ones obtained in the reference experiment using original (un-encoded) speech.

The FR standard turns out to be much more sensitive to errors than HR in both tasks. Specifically, the transmission errors significantly impoverish the recognition performance for BER higher than 10^{-4}. For the IDR task the losses are between 4.6%, for 10^{-3}, and 15.1% for 10^{-2} (we have not considered 5.10^{-2}, since in this case the word error rate is extremely low). While for CSR task, the losses are even higher, reaching 51.6% for a BER of 10^{-2}.

On the contrary, the HR is much more robust against errors and it is necessary to look at a BER of 5.10^{-2} in the IDR task or 10^{-2} in the CSR one to find significant recognition losses.

2) Tandeming: To reduce the number of experiments, in this case the G.726 stage is not included and both encoding stages (HR or FR) have been contaminated using the same channel conditions.

Fig. 8 shows the results for both IDR and CSR tasks. As expected, the recognition figures decrease notably with respect to an ideal situation. As it can be observed, those configurations including a FR stage are more severely affected, due to the higher sensitivity to errors of this codec. As an illustrative example, the recognition accuracy in the CSR task descends to 15.61% (from 86.63%, without transmission errors) for a FR-FR configuration and a BER of 5.10^{-3}. On the contrary, the recognition figure does not decrease at all for a HR-HR configuration and the same BER.

D. Evaluation of the Energy Estimation Procedure

In this subsection, we investigate the performance of the energy estimation procedure described in Section IV.B.3 for the CSR task. The training is performed on 'clean' (not affected by transmission errors) speech.

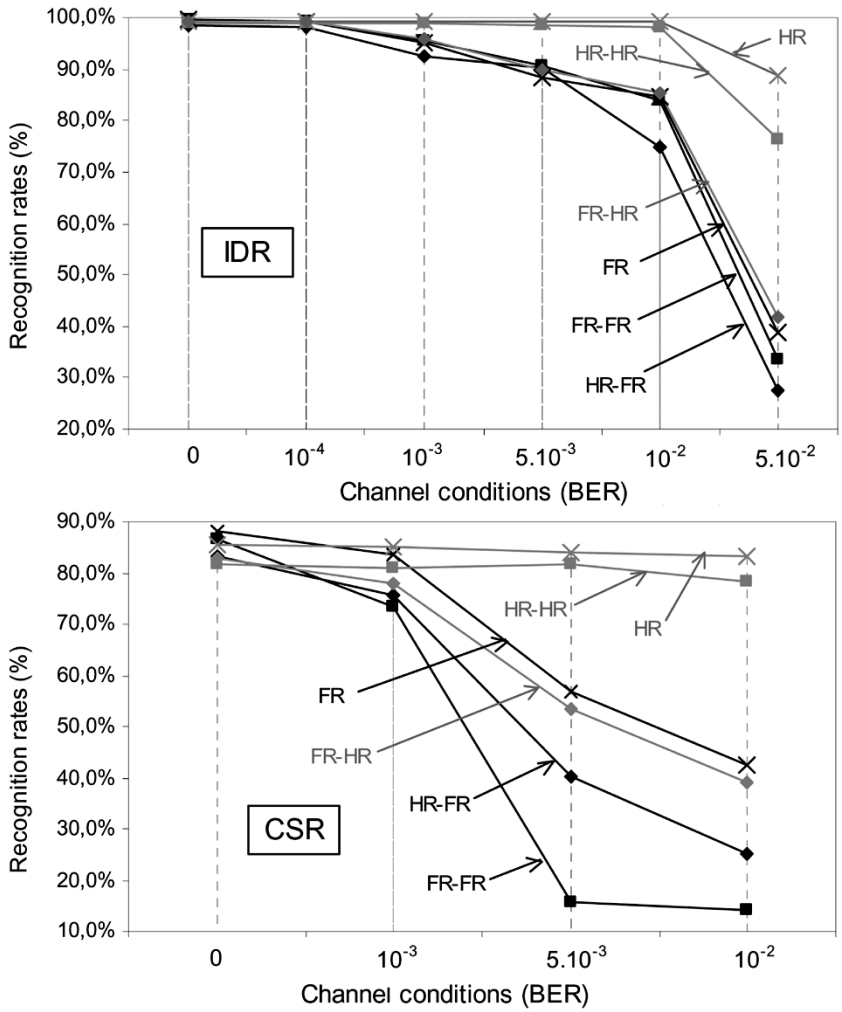

Fig. 8. Influence of transmission errors and tandeming encodings on the recognition performance for the IDR and CSR tasks. Several tandeming configurations are considered involving the HR and FR codecs. Also, channel conditions corresponding to several theoretical BERs are considered. As it can be seen, those configurations including a FR stage are more severely affected, due to the higher sensitivity to errors of this codec.

Fig. 9 compares the results obtained for the two methods of estimating the energy: 1) from the decoded speech; and 2) for the proposed method. The estimation procedure turns out to be more robust against transmission errors for BER higher than 10^{-3} (improvements of 6.2% and 11.1% can be observed for the $5{\cdot}10^{-3}$ and 10^{-2}, respectively). Below this BER the differences are not statistically significant.

E. Recognition From GSM Digital Speech

Along this subsection we compare the performances achieved by the proposed front-end with those obtained by the conventional one, for the two tasks at hand. We have labeled our front-end as "GSM Digital" while the conventional one as "Decoded Speech." The performances of both systems have been evaluated for the four (or five, depending on the case) channels described in the Table III. In any case, the training is performed on "clean" (not affected by transmission errors) speech.

As in previous subsections we will first treat each codec separately and then consider tandeming configurations.

1) One Coding Stage: The upper part of Fig. 10 shows the results for the IDR task. As can be observed, there are no significant differences between both front-ends when the HR is considered. In all likelihood, the task is too simple and therefore the performance is almost equally high irrespective of the accuracy of the parameterization employed. On the contrary, the proposed front-end is clearly superior for the FR case and BERs higher than 10^{-4}. Furthermore, the improvement obtained is higher as the channel conditions are poorer, going from 3% for a BER of 10^{-3} up to 15.6% for 5.10^{-2}.

The results for the CSR task are displayed in lower part of Fig. 10. In this case, the suggested front-end is substantially superior to the conventional one for the HR, with the improvement extending from 3.2% for a BER of 0% to 26.6% for 5.10^{-2}. The result obtained in clean conditions $(\mathrm{BER} = 0)$ clearly indicates that the proposed front-end is effective in avoiding the coding distortion. Besides, this improvement increases with the BER. Therefore, the front-end is also effective against transmission errors.

For the FR the results are not so good. In fact, significant differences only appear for BER higher than 10^{-3}. In our opinion, in this case the front-end is not good at coping with the coding distortion because the digital representation of the speech spectral envelope in the FR codec is not as precise as in the HR (The FR only extracts and codes 8 LP parameters, while the

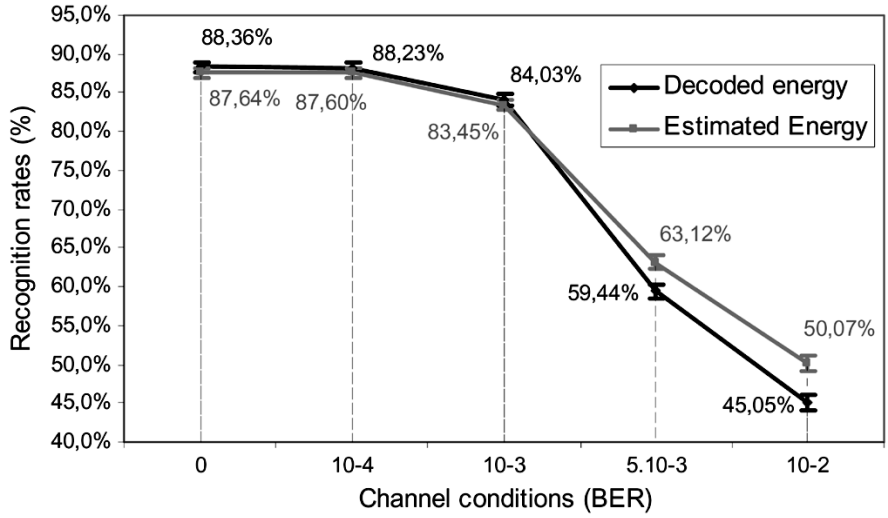

Fig. 9. Performance comparison of the proposed GSM Digital front-end using two different values for the energy. The solid line corresponds to the energy obtained from the decoded speech, while the dashed line is for the energy estimation method described in Section IV.B.3. Several channel conditions are considered, corresponding to theoretical BERs ranging from 10^{-4} to 5.10^{-2}. The proposed estimation procedure turns out to be more robust for BERs higher than 10^{-3}. Below this BER, the differences are not statistically significant.

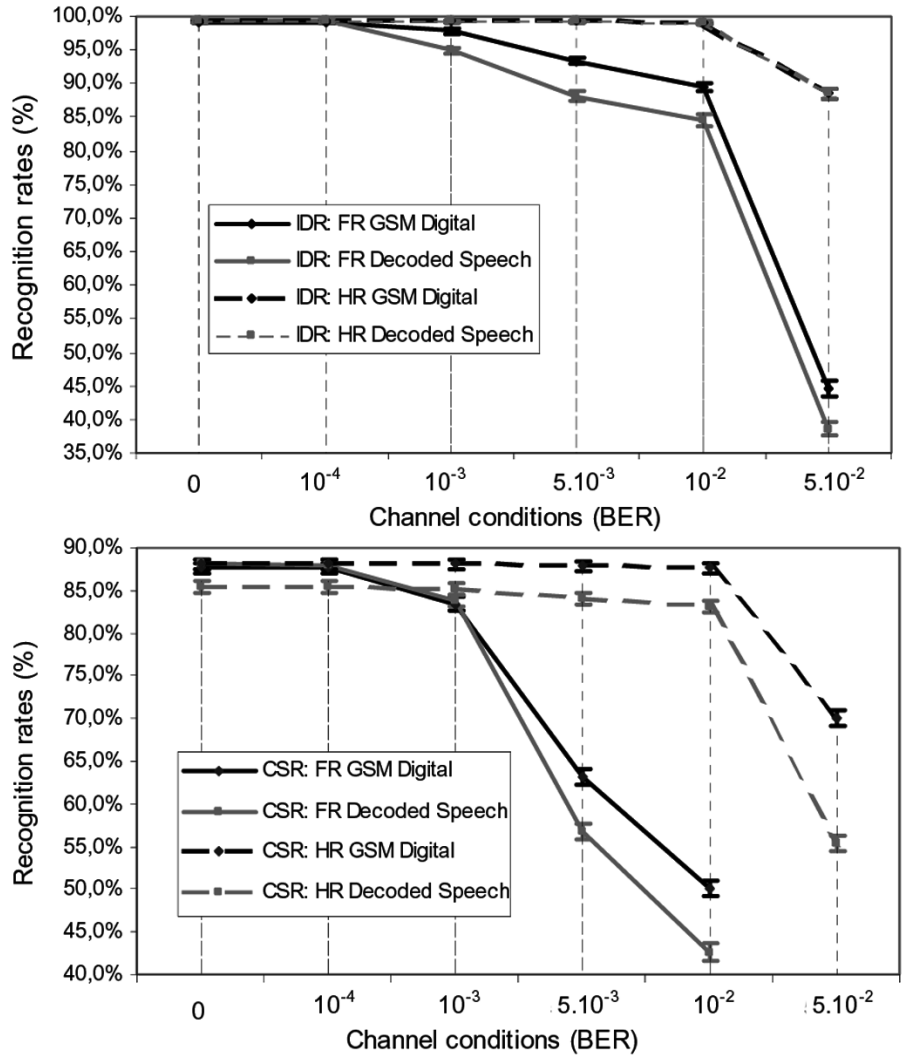

Fig. 10. Performance comparison of the proposed front-end ('GSM Digital') with the conventional one ("Decoded Speech") for the IDR and CSR tasks. Several channel conditions are considered, corresponding to theoretical BERs ranging from 10^{-4} to 5.10^{-2}.

TABLE I
COMPARISON PERFORMANCE OF THE PROPOSED (LOWER RESULT) AND CONVENTIONAL (UPPER ONE) FRONT-ENDS FOR THE IDR TASK AND FOR SEVERAL TANDEMING CONFIGURATIONS INVOLVING THE HR AND THE FR CODECS. TRANSMISSION ERRORS ARE ALSO CONSIDERED. BOLDFACE FONT IS USED TO HIGHLIGHT STATISTICALLY SIGNIFICANT IMPROVEMENTS

IDR Task						
Tandem. Config.	Recognition Rate (%) (95% Confidence interval)					
	BER = 0	BER = 10^{-4}	BER = 10^{-3}	BER = 5.10^{-3}	BER = 10^{-2}	BER = 5.10^{-2}
---	99.66%	-	-	-	-	-
	99.66%					
FR	99.51%	99.39%	**94.94%**	**88.18%**	**84.53%**	**38.64%**
	99.26%	99.22%	**97.83%**	**93.43%**	**89.52%**	**44.66%**
FR-FR	**99.44%**	99.22%	**95.62%**	**90.58%**	**83.74%**	**33.52%**
	99.03%	99.00%	**97.06%**	**92.97%**	**88.95%**	**37.54%**
HR-FR	98.31%	**98.09%**	**92.66%**	**90.39%**	**74.67%**	**27.37%**
	98.80%	**98.72%**	**96.82%**	**96.82%**	**82.98%**	**37.89%**
HR	99.42%	99.42%	99.38%	99.29%	99.07%	88.56%
	99.39%	99.39%	99.38%	99.32%	99.04%	88.50%
HR-HR	98.74%	98.74%	98.78%	98.51%	98.04%	76.28%
	99.03%	99.03%	98.99%	98.84%	98.40%	77.10%
FR-HR	99.27%	99.22%	**95.69%**	**89.90%**	**85.21%**	**41.74%**
	99.28%	99.26%	**98.13%**	**94.95%**	**91.81%**	**49.19%**

HR works with 10). On the other hand, this result confirms the findings in [33] indicating that the residue contains information relevant to the recognition.

2) Tandeming: As in the previous subsection, both encoding stages (HR or FR) have been contaminated using the same channel conditions.

In this case, the proposed front-end will only be able to deal with the distortions arising from the last coding stage. Nevertheless, as shown by the results obtained, it turns out to be extremely effective.

Table I shows the results for the IDR task comparing the conventional front-end (upper result) with the proposed one (lower result). As can be observed, wherever a FR is involved the digital front-end produces significant improvements which go from 1.51% for the FR-FR configuration and a BER of 10^{-3} to 38.44% for HR-FR and 5.10^{-2}. It is worth noting that these configurations are the ones in which the improvement margin is bigger due to the inferior behavior of the FR codec under channel distortions. For the HR-HR configuration, however, both front-ends are equivalent. As discussed in the case of one coding stage case, the task is too simple to make significant differences.

Similarly, the results for the CSR task are shown in Table II. Again, as in the case of one coding stage, the digital front-end is substantially superior to the conventional one when a HR stage is involved. The reason is two-fold: the spectral envelope is precisely represented in the HR, and now the task is complex enough to be sensitive to the more precise GSM digital parameterization. For example, the improvement, for a HR-HR configuration increases from 3.3% for a BER of 0 to 5.4% for 10^{-2}. Also, for a FR-HR configuration, the improvement goes from 4.6% for a BER of 0 to 52.2% for 10^{-2}.

VI. CONCLUSIONS AND FURTHER WORK

After reviewing the difficulties that speech recognition technologies face in a digital mobile phone environment, we have proposed a new front-end for speech recognition on mobile networks. In particular, we suggest performing the recognition from the encoded speech (i.e., the bitstream). In this way, we are circumventing the influence of some sources of distortion pertaining the encoding–decoding process. Furthermore, when transmission errors occur, our front-end is not affected by errors in bits encoding excitation parameters.

In the first place, we have evaluated the influence of coding distortion (considering several tandeming configurations) and transmission errors on ASR tasks. Our experiments have revealed, in accordance with some previous works ([23], [10], [42], [26], [27]), that the coding distortion can severely affect the recognition figures. Specifically, higher recognition rates are achieved for FR than for HR, as expected taking into account their respective bit rates. Besides, the more complex the ASR task is, the bigger the recognition losses become. Finally, the common (in practice) tandeming connections dramatically worsen the situation.

As regards the transmission errors, leaving aside the obvious observation that the higher the BER is, the poorer the recognition figures become, we can draw a different conclusion

TABLE II
COMPARISON PERFORMANCE OF THE PROPOSED (LOWER RESULT) AND CONVENTIONAL (UPPER ONE) FRONT-ENDS FOR THE CSR TASK AND FOR SEVERAL TANDEMING CONFIGURATIONS INVOLVING THE HR AND THE FR CODECS. TRANSMISSION ERRORS ARE ALSO CONSIDERED. BOLDFACE FONT IS USED TO HIGHLIGHT STATISTICALLY SIGNIFICANT IMPROVEMENTS

CSR Task				
Tandem. Config.	Recognition Rate (%) (95% Confidence interval)			
	BER = 0	BER = 10^{-3}	BER = 5.10^{-3}	BER = 10^{-2}
---	90.83%	-	-	-
FR	88.10%	83.74%	**56.79%**	**42.62%**
	87.64%	83.45%	**63.12%**	**50.07%**
FR-FR	86.63%	**73.51%**	**15.61%**	14.14%
	85.58%	**75.33%**	**21.84%**	14.11%
HR-FR	83.14%	75.62%	**40.01%**	**24.96%**
	82.10%	75.19%	**46.32%**	**30.93%**
HR	**85.39%**	**85.28%**	**84.11%**	**83.10%**
	88.09%	**88.10%**	**87.86%**	**87.62%**
HR-HR	**81.53%**	**81.12%**	**81.54%**	**78.27%**
	84.22%	**84.38%**	**84.55%**	**82.46%**
FR-HR	**82.98%**	**77.83%**	**53.54%**	**38.87%**
	86.82%	**83.53%**	**68.07%**	**59.17%**

depending on the coder being tested. In particular, the HR has turned out to be much more robust than the FR. This fact becomes evident in both one coding and tandeming scenarios (the tandem configurations including a FR stage become less robust).

In the second place, we have evaluated our front-end and compared it to the conventional approach in two ASR tasks, namely, speaker-independent IDR, and speaker-independent CSR. The comparison includes tandeming configurations and has been conducted in several simulated channel conditions derived from a channel model briefly summarized in the Appendix.

Again, the conclusions are different for each codec. In the case of HR, and for the CSR task, the digital front end is notably superior (with improvements of up to 26.6%) to the conventional one for any of the considered conditions (for one coding stage or tandeming connections and from error free to the highest BER).

In the case of the FR, and also for the CSR task, the conclusions are not so optimistic. The suggested front-end becomes equivalent to the conventional for clean conditions and low BERs. Nevertheless, as the channel conditions worsen, the proposed front-end becomes more effective.

As a general conclusion, it can be said that the suggested front-end is clearly effective in coping with transmission errors, since the achieved improvement becomes more important when the BER increases. Furthermore, if the speech spectral envelope is precisely encoded (as in the HR), the front-end is also highly effective against coding distortion. Besides, the FR codec is being progressively substituted by the Enhanced Full Rate (EFR) which is compatible to the proposed front-end parameterization and encodes the spectral envelope as precisely as HR.

Thus, though this paper has focused on the European GSM system and on the HR and FR speech codecs, this approach can be applied to other speech codecs in GSM (e.g., EFR or AMR), since, in every case, low bit-rate codecs typically used in cellular systems are CELP-type and, consequently, encode and transmit the spectral envelope of the speech signal. Besides, our main conclusions could be extended to other digital cellular systems (as shown in [37]).

Currently, we are working on the extension of our experiments to the EFR and AMR codecs. In both cases, we deal with an ACELP-type algorithm. Furthermore, EFR is identical to AMR at 12.2 kb/s. The GSM digital approach needs to extract two different kinds of information from the bitstream: the spectral envelope and an energy value. The spectral envelope is represented by LSP parameters. We have already studied the LSP to LP cepstral coefficients conversion for another ACELP codec (at 5.3 kb/s), namely: G.723.1 [45]. Furthermore, Kim *et al.* proposed a straightforward transformation of LSP into LP cepstral coefficients (pseudocepstrum) [37]. With respect to the energy estimation, the procedure to follow is very similar to that already developed for G.723.1 [45].

The combined influence of additive noise and the distortions considered in this paper is another interesting research line for future work. The noise addition contributes significantly to the encoding-decoding process distortion, since the signal to be encoded goes away from the speech model considered

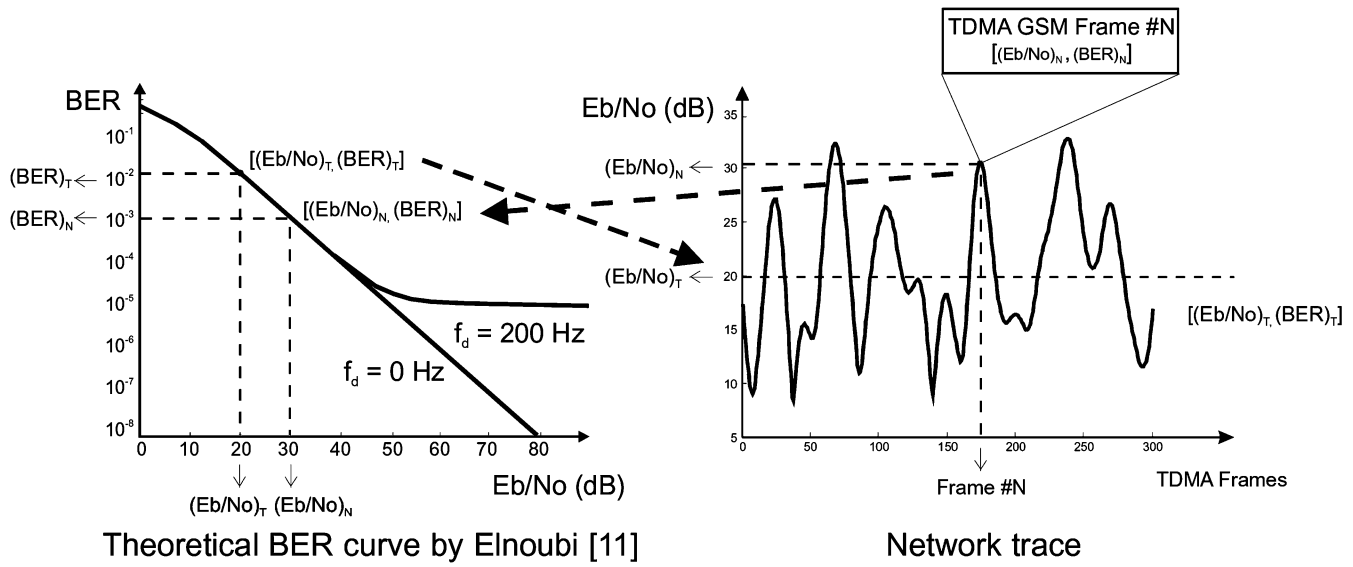

Fig. 11. This figure describes the procedure used for introducing bursty errors in the TDMA GSM frames. The value of the target signal-to-noise ratio, $(E_b/N_o)_T$, corresponding to the target BER, $(\mathrm{BER})_T$, is determined by using the theoretical curve (on the left) derived by Elnoubi [11]. This curve represents the BER vs. the signal-to-noise ratio (E_b/N_o) due to the short-term variations of the received signal (Rayleigh fading). The network trace (on the right) accounts for slow variations (shadowing) around the value of $(E_b/N_o)_T$ responsible for bursty errors. This trace is generated from a lognormal distribution with a sampling rate equal to the TDMA frame rate.

by the encoder. Therefore, it is expected that the conventional approach is more sensitive than the digital front-end to this type of distortion. Nevertheless, this should be researched, since the parameterization extracted from the bitstream is also affected by noise.

The discontinuous transmission (DTX) is another component of a digital mobile phone system that deserves to be researched due to its presumable influence on ASR.

Finally, in our opinion, there is still room to investigate more sophisticated ways of coping with transmission errors and frame erasures, within the framework we have described in this paper. For example, a specific concealment techniques could be designed to cope with frame erasures.

APPENDIX
GSM CHANNEL SIMULATION PROCEDURE

This Appendix briefly describes the GSM channel simulation procedure used in this paper. First, we will describe the characteristics of the GSM channel and the methodology used to generate realistic network traces (signal-to-noise ratio vs. time). Second, we will describe how we have used these traces to simulate the insertion of bursty transmission errors in the bitstream. More details about this procedure can be found in [28].

In mobile communications, the most destructive source of distortion is the fading phenomena [48] which makes transmission errors not to be randomly distributed, but present a bursty behavior. As a result of the fading effect, the received signal usually consists of the sum of a number of (attenuated) copies of the transmitted signal with different phases. As it is not possible to know the amplitude and phase of these multipath components, it is necessary to model the mobile propagation channel in a statistical way.

In order to simplify the analysis, the wireless channel is supposed to be affected by two overlapping random processes with different temporal characteristics [5]. The first one is the *slow* or long-term random process (*shadowing*), which is related to large scale signal amplitude variations due to the presence of obstacles like buildings in urban areas. In this process, it is typically assumed that the signal-to-noise ratio of the received signal follows a lognormal distribution. The second one is the *short-term* random process (*Rayleigh fading*), which is mainly related to the mobile speed, and it produces fluctuations in the received signal over small time periods. In this case, a Rayleigh distribution is typically assumed. In brief, the signal arrives at the receiver with a signal-to-noise ratio (E_b/N_o) that suffers important variations as a function of time. When E_b/N_o is small, the BER increases and the probability of appearance of a fading interval is high. A representation of these variations (E_b/N_o vs. time) is called a network trace and an example of it is shown in the right side of Fig. 11 (other details depicted in Fig. 11 will be explained further on).

In our previous works [26], [27], the network modeling approach used was based on a Gilbert model (i.e., two state discrete Markov chain) [35]. However, this model is too simple and it cannot adequately capture the nature of the actual GSM traces [40]. The two fold stochastic model described above considerably increases the realism of the simulated channel. This approach has been previously used by Wigard *et al.* [56] for UMTS link level simulation and we have adapted it for GSM channel simulation. Our method relies on the combination of available theoretical results (for modeling the Rayleigh fading phenomena) and measured data (for shadowing effects). The key feature of this model is the use of a closed-form expression for the mean (theoretical) BER that characterizes the *short-term* channel behavior. We have used the expression derived by Elnoubi [11] for a fast Rayleigh fading channel. This expression also accounts for effect of the GMSK modulator [15] and different mobile speeds. The BER curve shown in the left side of Fig. 11 is a representation of this analytical expression

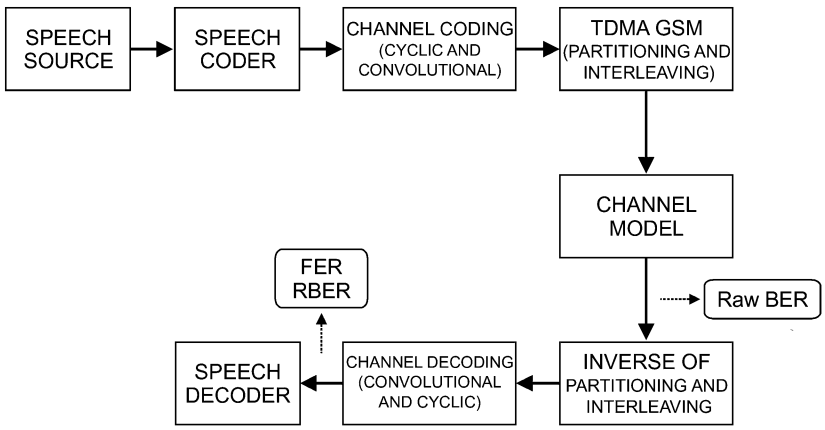

Fig. 12. Top-level block diagram of the GSM channel simulator. The overall system includes implementations of the channel model (see Fig. 11), the speech codecs and channel coding (cyclic, convolutional coding) according to the ETSI/GSM specifications for both full-rate and half-rate speech traffic channels. Also, the blocks relevant to the arrangement of the digital TDMA stream (reordering, partitioning, interleaving and burst formatting) were implemented. Raw BER (BER before channel decoding and de-interleaving) represents the percentage of errors introduced by the channel in the bitstream. The values of FER ("Frame Error Rate") and RBER ("Residual Bit Error Rate") indicate, respectively, the percentage of erroneous frames and the remaining transmission errors appearing at the input of the speech decoder.

TABLE III
CHARACTERISTICS OF THE GSM CHANNELS USED FOR BOTH, FULL-RATE AND HALF-RATE. THEORETICAL BER, RAW BER, FER ('FRAME ERASURE RATE') AND RBER ('RESIDUAL BIT ERROR RATE') VALUES ARE SHOWN FOR EACH CHANNEL. RAW BER, FER, AND RBER ARE NOT THE THEORETICAL VALUES, BUT THOSE EXPERIMENTALLY COMPUTED OVER THE DATABASES USED

	Full-rate				Half-rate			
Channel	Th. BER	Raw BER	FER	RBER	Th. BER	Raw BER	FER	RBER
Channel0	0	0%	0%	0%	0	0%	0%	0%
Channel1	10^{-4}	0.0111%	0%	0.0014%	10^{-4}	0.0131%	0%	0.0010%
Channel2	10^{-3}	0.1601%	0.019%	0.0372%	10^{-3}	0.1635%	0.015%	0.0265%
Channel3	5.10^{-3}	1.0371%	0.405%	0.3199%	5.10^{-3}	1.0672%	0.176%	0.1622%
Channel4	10^{-2}	1.6118%	1.130%	0.5033%	10^{-2}	1.6465%	0.479%	0.2753%
Channel5	5.10^{-2}	-	-	-	5.10^{-2}	7.0999%	12.333%	2.3222%

for two different Doppler frequency shifts f_d, namely, 0 Hz and 200 Hz. The *long-term* variations are represented by a lognormal distribution whose parameters have been determined upon the statistical analysis over collected network traffic traces.

Fig. 11 illustrates the process for the generation of GSM network traces (and the corresponding transmission errors) for a certain mean BER we will call *target BER*, $(\mathrm{BER})_T$. This process, depicted in the *Channel Model* block diagram of Fig. 12, involves the following steps:

- Reading of the target signal-to-noise ratio, $(E_b/N_o)_T$ corresponding to the considered $(\mathrm{BER})_T$ from the theoretical BER curve.
- Generation of a network trace from a lognormal distribution for a given sampling rate. This trace represents how the actual E_b/N_o value for a given frame N moves around $(E_b/N_o)_T$ producing bursty errors. Here, the sampling rate is equal to the TDMA frame rate in GSM [18].
- Mapping of the $(E_b/N_o)_N$ into its corresponding $(\mathrm{BER})_N$ using the theoretical BER curve (left side of Fig. 11). Finally, each bit is either contaminated or not according to $(\mathrm{BER})_N$.

Therefore, at the output of the channel modeling block we measure what we call the *Raw BER* (BER before decoding and de-interleaving) as the number of erroneous bits averaged over all the frames.

Following the block diagram of Fig. 12, the bitstream undergoes the inverse partitioning and interleaving processes and reaches the channel decoder which is able to detect and correct some of the errors. Finally, the uncorrected errors can be

grouped into two classes: frame erasures and residual bit errors. For our experiments, we have characterized these two types of errors by the following:

- *Frame Erasure Rate (FER)*: It measures the number of erroneous frames that were replaced by a concealing mechanism.
- *Residual Bit Error Rate (RBER)*: It measures the remaining transmission errors, which were not corrected or detected in the channel decoding stage.

From the speech recognizer point of view, FER and RBER are the parameters that define the GSM scenario because they are the sources of distortion that actually affect the speech (source) decoding. These parameters vary enormously according to the propagation conditions, which are mainly characterized by the mobile speed and the particular scenario: rural area, hilly terrain, urban area, etc. Maximum values for FER and RBER can be found in the ETSI Recommendation [16] for different propagation conditions. The simulated channels are realistic in the sense that they meet the reference performance criteria defined by the ETSI recommendation [16] for both, full-rate and half-rate speech traffic channels. For instance, the maximum FER and RBER values established by ETSI for a FR traffic channel with static conditions are around 0.1% and 0.77%, respectively. In our simulations, the FR channel with a BER target of 10^{-4} and a FER of 0.0111% and a RBER of 0.0014% (see Table III) can be considered a static channel. In the same way, simulated FR channels with a BER target below 5.10^{-3} correspond to different propagation conditions in an urban scenario considering a mobile speed of 50 km/h. In this latter case, the maximum FER and RBER values allowed in the ETSI recommendation are around 3% and 0.51% respectively. We have not considered channels with a BER value below 5.10^{-2} (this minimum value is set to 10^{-2} in CSR task for the FR codec) because the corresponding word error rate becomes very and it would be difficult to extract useful conclusions.

Therefore, in our experimentation, we have considered a realistic and complete GSM scenario which includes not only a channel model but also the GSM channel coding/decoding processes. Fig. 12 shows the top-level block diagram of the whole system used in the experiments presented throughout this paper. Specifically, it includes implementations for the channel model described above, FR and HR speech codecs and their corresponding channel coding/decoding modules according to the ETSI/GSM specifications [14]. Also, the blocks relevant to the arrangement of the digital TDMA stream (reordering, partitioning, interleaving and burst formatting) specified in [13] were implemented. Both full-rate and half-rate speech traffic channels are considered.

Following this procedure, we have designed different GSM channels for both, full-rate and half-rate, whose characteristics are listed in Table III. Together with the theoretical BER (BER target) for each channel, raw BER, FER ("Frame Erasure Rate"), and RBER ("Residual Bit Error Rate") values are shown. Raw BER, FER, and RBER are not theoretical values, but experimentally computed ones for the databases we have employed.

ACKNOWLEDGMENT

The authors would like to thank their colleague M. A. Vázquez-Castro, without whom this work would have been much less realistic. She conceived the GSM channel simulation method and provided the authors with the software for transmission error generation.

REFERENCES

[1] N. Benvenuto, G. Bertocci, and W. R. Daumer, "The 32-kb/s ADPCM coding standard," *AT&T Tech. J.*, vol. 65, pp. 12–22, Sep./Oct. 1986.

[2] S. F. Campos-Neto, F. L. Corcoran, and A. Karahisar, "Performance assessment of tandem connection of enhanced cellular coders," in *Proc. ICASSP'99*, vol. I, Phoenix, AZ, 1999, pp. 177–180.

[3] S. H. Choi, H. K. Kim, and H. S. Lee, "Speech recognition using quantized LSP parameters and their transformations in digital communication," *Speech Commun.*, vol. 30, pp. 223–233, 2000.

[4] G. Cohen, T. Ramabadran, and R. Tucker, "Requirements for speech reconstruction from the standard cepstral features," *Work Document STQ AURORA DSR WG*, Jan. 2001.

[5] G. D'Aria, F. Muratore, and V. Palestine, "Simulation and performance of the pan-european land mobile radio system," *IEEE Trans. Veh. Technol.*, vol. 41, no. 2, pp. 177–189, May 1992.

[6] J. de Veth, F. de Wet, B. Cranen, and L. Boves, "Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR," *Speech Commun.*, vol. 34, no. 1–2, pp. 57–74, 2001.

[7] P. A. Devijver and J. Kitter, *Pattern Recognition: A Statistical Approach*. London, U.K.: Prentice-Hall, 1982.

[8] V. V. Digalakis, L. G. Neumeyer, and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE J. Select. Areas Commun.*, vol. 17, no. 1, pp. 82–90, Jan. 1999.

[9] S. Dimolistsas, F. L. Corcoran, C. Ravishankar, and M. Baraniecki, "Voice quality of interconnected PCS, Japanese cellular, and public switched telephone networks," in *Proc. ICASSP'95*, vol. I, Detroit, MI, 1995, pp. 273–276.

[10] S. Dufour, C. Glorion, and P. Lockwood, "Evaluation of the root-normalized front-end (RN_LFCC) for speech recognition in wireless gsm network environments," in *ICASSP-96*, vol. I, Atlanta, GA, 1996, pp. 77–80.

[11] S. M. Elnoubi, "Analysis of GMSK with differential detection in land mobile radio channels," *IEEE Trans. Veh. Technol.*, vol. VT-35, no. 4, Nov. 1986.

[12] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithms*, Apr. 2000. ETSI Std. ES 201 108.

[13] *Digital Cellular Telecommunications System (Phase 2). Physical Layer on the Radio Path. General Description*, Sep. 1994. ETSI Recommend. GSM 05.01 (ETS 300 573).

[14] *Channel Coding*, Feb. 1992. ETSI Recommend. GSM 5.03.

[15] *Digital Cellular Telecommunications System (Phase 2). Modulation*, Feb. 1992. ETSI Recommend. GSM 05.04 (ETS 300 576).

[16] *Digital Cellular Telecommunications System (Phase 2+); Radio Transmission and Reception*, Jan. 1998. ETSI Recommend. GSM 05.05 ver. 5.7.0 (DRAFT prETS 30910).

[17] *Digital Cellular Telecommunications System; Full Rate Speech Transcoding*, Feb. 1992. ETSI Recommend. GSM 6.10.

[18] *Substitution and Muting of Lost Frames for Full Rate Speech Channels*, Feb. 1992. ETSI Recommend. GSM 6.11.

[19] *Digital Cellular Telecommunications Systems; Half Rate Speech; Part 2: Half Rate Speech Transcoding*, Dec. 1995. ETSI Recommend. GSM 6.20.

[20] *Substitution and Muting of Lost Frames for Half Rate Speech Traffic Channels*, Dec. 1995. ETSI Recommend. GSM 6.21.

[21] *Enhance Full Rate (EFR) Speech Transcoding*, Mar. 1997. ETSI Recommend. GSM 6.60.

[22] *Digital Cellular Telecommunication Systems: Adaptive Multi-Rate (AMR) Speech Transcoding*, Apr. 2000. ETSI Recommend. GSM 06.90.

[23] S. Euler and J. Zinke, "The influence of speech coding algorithms on automatic speech recognition," in *Proc. ICASSP'94*, vol. I, Adelaide, Australia, 1994, pp. 621–624.

[24] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, 2nd ed. New York: Marcel Dekker, 2001.

[25] M. J. F. Gales, ""Nice" model-based compensation schemes for robust speech recognition," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997, pp. 55–66.

[26] A. Gallardo-Antolín, F. Díaz-de-María, and F. Valverde-Albacete, "Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks," in *Proc. ICASSP'99*, vol. I, Phoenix, AZ, 1999, pp. 277–280.
[27] ——, "Recognition from GSM digital speech," in *Proc. ICSLP'98*, Sidney, NSW, Australia, 1998.
[28] A. Gallardo-Antolín, M. Vázquez-Castro, F. Díaz-de-María, F. Valverde-Albacete, and F. Pérez-Fontán, "BER performance assessment of the land mobile GSM channel with application to automatic speech recognition tasks," in *Proc. 5th Bayona Workshop on Emerging Technologies in Telecommunications*, vol. 1, 1999, pp. 212–216.
[29] I. A. Gerson and M. A. Jasiuk, "Vector sum excited linear prediction (VSELP)," in *Advances in Speech Coding*. Norwell, MA: Kluwer, 1991, pp. 69–79.
[30] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, 1995.
[31] T. Honkanen, J. Vainio, K. Järvinen, P. Haavisto, R. Salami, C. Laflame, and J.-P. Adoul, "Enhanced full rate speech codec for IS-136 digital cellular system," in *Proc. ICASSP'98*, vol. 1, Munich, Germany, 1998, pp. 731–734.
[32] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[33] J. M. Huerta and R. M. Stern, "Speech compression from GSM codec parameters," in *Proc. ICSLP'98*, vol. 4, Sidney, NSW, Australia, 1998, pp. 1463–1466.
[34] J. C. Junqua, *Robust Speech Recognition in Embedded Systems and PC Applications*. Norwell, MA: Kluwer, 2000.
[35] L. N. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proc. of the IEEE*, vol. 66, pp. 724–744, Jul. 1978.
[36] L. Karray, A. B. Jelloun, and C. Mokbel, "Solutions for robust recognition over the GSM cellular network," in *ICASSP-98*, vol. 1, Munich, Germany, 1998, pp. 261–246.
[37] H. K. Kim, S. H. Choi, and H. S. Lee, "On approximating line spectral frequencies to LPC cepstral coefficients," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 2, Mar. 2000.
[38] H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communication system," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 558–568, 2001.
[39] A. M. Kondoz, *Digital Speech: Coding for Low Rate Communication Systems*. New York: Wiley, 1996.
[40] A. Konrad, B. Y. Zhao, A. D. Joseph, and R. Ludwig, "A Markov-based channel model algorithm for wireless networks," in *Proc. 4th ACM Int. Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, 2001.
[41] C.-H. Lee, "On stochastic feature and model compensation approach to robust speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 29–47, Aug. 1998.
[42] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. ICSLP'96*, vol. IV, Philadelphia, PA, 1996, pp. 2344–2347.
[43] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova, "Toward improving ASR robustness for PSN and GSM telephone applications," *Speech Commun.*, vol. 23, pp. 141–159, 1997.
[44] C. Nadeu, P. Pachés-Leal, and B.-H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Commun.*, vol. 22, pp. 315–332, 1997.
[45] C. Peláez-Moreno, A. Gallardo-Antolín, and F. Díaz-de-María, "Recognizing voice over IP: A robust front-end for speech recognition on the world wide web," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 209–218, 2001.
[46] G. N. Ramaswamy and P. S. Gopalakrishnan, "Compression of acoustic features for speech recognition in network environments," in *Proc. ICASSP'98*, 1998.
[47] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.
[48] S. M. Redl, M. K. Weber, and M. W. Oliphant, *An Introduction to GSM*. Boston, MA: Artech House, 1995.
[49] T. Salonidis and V. Digalakis, "Robust speech recognition for multiple topological scenarios of the GSM mobile phone system," in *ICASSP-98*, vol. 1, Seattle, WA, 1998, pp. 101–104.
[50] K. Shen, J. Bin, and G. Cohen, "Standardization of pitch determination, compression and transmission for DSR terminals," in *Work Document STQ AURORA DSR WG*, Feb. 2001.
[51] R. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Automatic speech and speaker recognition," in *Signal Processing for Robust Speech Recognition*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Dordrecht, The Netherlands: Kluwer, 1996, ch. 15.
[52] *NIST Resource Management Corpus (RM1)*, 1992.
[53] R. Tucker, T. Robinson, J. Christie, and C. Seymour, "Compression of acoustic features—Are perceptual quality and recognition performance incompatible goals?," in *Proc. Eurospeech'99*, vol. 5, Budapest, Hungary, 1999, pp. 2155–2158.
[54] P. Vary, R. Hoffman, K. Hellwig, and R. Sluyter, "A regular-pulse excited linear predictive coder," *Speech Commun.*, vol. 7, no. 2, pp. 209–215, 1988.
[55] N. A. Weiss and M. J. Hasset, *Introductory Statistics*, 3rd ed. Reading, MA: Addison-Wesley, 1993.
[56] J. Wigard and P. Mogensen, "A simple mapping from C/I to FER and BER for a GSM type of air interface," in *Proc. PIMRC'96*, Taipei, Taiwan, R.O.C., Oct. 1996, pp. 78–82.
[57] S. Young *et al.*, *HTK-Hidden Markov Model Toolkit (Ver. 2.1)*. Cambridge, MA: Cambridge Univ., 1995.
[58] F. Xie and D. Van Compernolle, "Speech enhancement by spectral magnitude estimation—A unifying approach," *Speech Commun.*, vol. 19, pp. 89–104, 1996.

Ascensión Gallardo-Antolín received the telecommunication engineering degree and the Ph.D. degree from the Polytechnic University of Madrid, Madrid, Spain, in 1993 and 2002, respectively.

Since 1997 she has been Assistant Professor at the Universidad Carlos III, Madrid. Her research interests are in speech recognition, spoken information retrieval, and dialog and signal processing for multimedia human–machine interaction. She has coauthored several communications to international conferences, mainly in speech recognition. She has participated in several research projects including some of the Spanish Council on Science and Technology and the UE.

Carmen Peláez-Moreno (M'03) received the Telecommunication Engineering degree from the Public University of Navarre, Pamplona, Spain, in 1997 having completed her final year project in Westminster University, London, U.K. She obtained the Ph.D. degree from the University Carlos III, Madrid, Spain, in 2002.

Since 2001, she has been Assistant Professor with the University Carlos III and is currently on sabbatical leave as Research Visitor with the International Computer Science Institute, Berkeley, CA.

Dr. Peláez-Moreno received the Best Ph.D. Thesis Award from the Spanish Telecommunication Engineers Association in 2002.

Fernando Díaz-de-María (M'97) received the Telecommunication Engineering degree in 1991 and the Ph.D. degree in 1996 from the Polytechnic University of Madrid, Madrid, Spain.

He is an Associate Professor at the Department of Signal Theory and Communications, University Carlos III, Madrid, Spain. He is currently Sub-Director of the Ph.D. Program for Telecommunication Technologies, Universidad Carlos III de Madrid. His primary research interests include robust speech recognition, nonlinear speech processing, video coding, and multimedia.